# Evaluator interview: grantmaking judgment and research context

September 17, 2026

## Contents


**Decode the attributed role brief and extract its explicit and implicit evaluation criteria.** **12**
Question . . . 12
Background . . . 12
Sub-Question 1: **Source Status and Epistemic License** . . . 12
Sub-Question 2: **Explicit Criteria and Likely Interview Evidence** 14
Sub-Question 3: **Implicit Criteria from Institutional Context** . . 19
Sub-Question 4: **Gaps, Alternative Readings, and Preparation Conversion** . . . 23

**Map Coefficient Giving's structure, funds, giving scale and strategic posture.** **28**
Question . . . 28
Background . . . 28
Sub-Question 1: **Fund Architecture and Giving Scale** . . . 29
Sub-Question 2: **Strategic Posture and Evaluator Implications** . 34

**Analyze the stated AI and biosecurity scaling rationale, urgency and short-timeline strategy.** **38**
Question . . . 38
Background . . . 38
Sub-Question 1: **1. Decompose the rationale: two drivers and one enabling constraint** . . . 39
Sub-Question 2: **2. Bigger, faster, proactive: one capacity strategy in three faces** . . . 43
Sub-Question 3: **3. Short timelines, hedged portfolios, and a posture that looks like a reversal** . . . 46
Sub-Question 4: **4. Biosecurity: a parallel rationale with a different evidentiary base** . . . 47

Sub-Question 5: **5. Implications for grantmaker priorities, with calibrated answers** . . . 49
Sub-Question 6: **6. Falsifiers, discounts on headline numbers, and how a grantmaker should update** . . . 51

**Reconstruct the day-to-day grantmaker workflow from investigation to approval.** **55**
Question . . . 55
Background . . . 55
Sub-Question 1: Entry Points: How an Opportunity Enters the Pipeline 56
Sub-Question 2: The Investigation Itself: Proportionate Effort and the Evidence That Would Change Your Mind . . . 59
Sub-Question 3: Write-Up, Review and Approval: Where the Bottlenecks Actually Sit . . . 62
Sub-Question 4: After the Yes: Relationship Management, the Loop Back, and a Proposed Stop Rule . . . 67

**Map CG's donor, advisory and partnership context relevant to evaluator work.** **71**
Question . . . 71
Background . . . 71
Sub-Question 1: Stakeholder Map: Money, Advice, Trust, and Conflicts 71

**Extract hiring signals and cultural expectations from the role brief and CG and aligned-actor posts.** **80**
Question . . . 80
Background . . . 80
Sub-Question 1: **Extract Hiring Signals and Cultural Expectations** . . . 80

**Internalize reasoning-transparency standards and practice applying them to grant memos.** **87**
Question . . . 87
Background . . . 87
Sub-Question 1: **A reasoning-transparency checklist for grant memos** . . . 88
Sub-Question 2: **Rewritten grant-memo paragraph: before, after, and the decision difference** . . . 93
Sub-Question 3: **Calibration drill: rewriting vague claims into transparent claims** . . . 98

**Build a worked cost-effectiveness, BOTEC, expected-value and social-ROI comparison.** **104**
Question . . . 104
Background . . . 104
Sub-Question 1: Selecting the Numeraire and the Counterfactual . . . 104

Sub-Question 2: Candidate Grant A: Effective Careers, $500,000 . . . 107
Sub-Question 3: Candidate Grant B: Bed-Net-Style Global Health, $500,000 . . . 110
Sub-Question 4: Stress-Testing and Deciding . . . 111

**Analyze hits-based giving, power laws and portfolio construction. 116**
Question . . . 116
Background . . . 116
Sub-Question 1: What Hits-Based Giving Asserts, and What It Does Not . . . 117
Sub-Question 2: The Expected-Value Mathematics and Why Sample Means Mislead . . . 119
Sub-Question 3: Selection Risk in Retrospective Hit Lists . . . 121
Sub-Question 4: Grant Size, Portfolio Shape, Follow-On Funding: The Memo . . . 124

**Master grant-program design across money, taste, dealflow, hustle and trust. 129**
Question . . . 129
Background . . . 129
Sub-Question 1: The five components as one system, not a checklist . 130
Sub-Question 2: The diagnostic template: finding the binding component and the cheapest fix . . . 132
Sub-Question 3: Worked comparison: Coefficient Giving versus Manifund regranting . . . 135
Sub-Question 4: Live disagreements and the evidence that would settle them . . . 138

**Set proportionate due diligence, evidence thresholds and investigation tiers. 142**
Question . . . 142
Background . . . 142
Sub-Question 1: **The Proportionality Principle: What Varies and What Does Not** . . . 142
Sub-Question 2: **Tiered Due-Diligence Checklist Keyed to Size, Risk, and Uncertainty** . . . 144
Sub-Question 3: **Red Flags, Decisive-Evidence Tests, and the Stop Rule** . . . 151
Sub-Question 4: **Worked Drill: $10m Multi-Year Grant, $5k Rapid Grant, and $200k-$2m First-Time Grant** . . . 165
Sub-Question 5: **Calibrated Interview Answer and Rehearsal** . . 166

**Learn grant-memo writing and WBS, CBS, OBS and RACI budget construction. 169**
Question . . . 169
Background . . . 169

Sub-Question 1: **Grant-memo architecture: the six moves, the three requirements, and proportionate diligence** . . . . . 170
Sub-Question 2: **Work breakdown structure: from theory of change to packages, schedule, milestones and tiers** . . . . 174
Sub-Question 3: **Cost breakdown structure: cost pools, single drivers, coverage, and justification** . . . . . 178
Sub-Question 4: **OBS, control accounts and RACI: putting an owner on every deliverable** . . . . . 182
Sub-Question 5: **Worked example: a tiered first-year budget with cost per outcome and a named funder bar** . . . . . 185
Sub-Question 6: **Where the method is informal, uncertain or misleading, and how to disclose it** . . . . . 193

**Map fund intermediaries, regranting and small-grant ecosystems.** **199**
Question . . . . . 199
Background . . . . . 199
Sub-Question 1: **The topology: what is an intermediary of what** 199
Sub-Question 2: **Scale bands, the small-check gateway, and why the market has seams** . . . . . 203
Sub-Question 3: **Trust, taste, dealflow, hustle and money: the central disagreement** . . . . . 205
Sub-Question 4: **Routing rules: when to recommend each vehicle** 207
Sub-Question 5: **Matching, credit allocation, and why impact does not add up** . . . . . 211
Sub-Question 6: **Failure modes, caveats, and what would change this map** . . . . . 212

**Practice judgment under uncertainty and generate alternative defensible answers.** **215**
Question . . . . . 215
Background . . . . . 215
Sub-Question 1: **Drill 1: Big initial bet or staged pilot for a high-conviction new organization?** . . . . . 216
Sub-Question 2: **Drill 2: Fund a regranting intermediary or build direct grantmaking?** . . . . . 218
Sub-Question 3: **Drill 3: Fund technically strong but politically infeasible governance work?** . . . . . 221

**Build a threat-model matrix for AI risk pathways.** **225**
Question . . . . . 225
Background . . . . . 225
Sub-Question 1: **Skeleton of the matrix: four pathway families, six columns, one overlap trap** . . . . . 226
Sub-Question 2: **Evidence versus speculation: what would move each estimate** . . . . . 231

Sub-Question 3: **Who disputes which pathway, and how the disagreement is structured** . . . 234
Sub-Question 4: **From matrix to memo: a worked application with two defensible answers** . . . 236

**Compare AI progress, timelines, takeoff speeds and forecasting evidence.** **239**
Question . . . 239
Background . . . 239
Sub-Question 1: Decomposing Timelines into Four Parameters and One Comparable Object . . . 240
Sub-Question 2: Classifying the Disagreements: Capability, Diffusion, Magnitude . . . 242
Sub-Question 3: Auditing the Forecasting Evidence: What Deserves Weight and How Each Signal Fails . . . 247
Sub-Question 4: Mapping Which Grant Decisions Move Under Short Versus Long Timelines . . . 250

**Map technical AI safety agendas and empirical safety research.** **256**
Question . . . 256
Background . . . 256
Sub-Question 1: What actually distinguishes one technical AI safety agenda from another? . . . 257
Sub-Question 2: What is the fundable bottleneck for each agenda, and what evidence establishes it as binding? . . . 259
Sub-Question 3: How do funder priorities and the tractability dispute reshape the map? . . . 263

**Map the AI governance and policy landscape from research to enforcement.** **267**
Question . . . 267
Background . . . 267
Sub-Question 1: **Trace the AI Governance Pipeline and Locate Grantmaker-Relevant Leverage** . . . 268

**Map key AI safety and security organizations, funders and field-building infrastructure.** **276**
Question . . . 276
Background . . . 276
Sub-Question 1: **Mapping axes: from categorical labels to functional roles** . . . 276
Sub-Question 2: **Funders and funding instruments: ticket bands, decision speed, and trust gates** . . . 280
Sub-Question 3: **Field-building and talent infrastructure: the reflexive layer** . . . 283

Sub-Question 4: **Research, evaluation, and policy organizations: producers, evaluators, and standard-setters** . . . 286
Sub-Question 5: **Evidence gaps, disputes, and maintenance rules for the map** . . . 289

**Analyze AI safety community norms, context and strategic disagreements.** **292**
Question . . . 292
Background . . . 292
Sub-Question 1: What context actually is, and why it cannot simply be taught . . . 293
Sub-Question 2: Community norms: what they do, and which ones survive growth . . . 300
Sub-Question 3: The disagreement map, with cruxes preserved . . . 308
Sub-Question 4: The earliest point of disagreement: a worked exercise with two defensible answers . . . 316

**Build a biosecurity and pandemic-preparedness brief.** **324**
Question . . . 324
Background . . . 324
Sub-Question 1: **Biological Risk Pathways and the Distinctions an Evaluator Must Preserve** . . . 325
Sub-Question 2: **CG's Biosecurity Strategy Compared with AI Safety and Global Health** . . . 328
Sub-Question 3: **A Biosecurity Grant Evaluation Rubric and Three Worked Examples** . . . 329
Sub-Question 4: **Disputed Estimates and How to Represent Them** 331
Sub-Question 5: **Generalist Comparison Across Biosecurity, AI Safety, and Global Health** . . . 334
Sub-Question 6: **Talent Bottlenecks and Proportionate Due Diligence: The ERA AIxBio Case** . . . 336

**Build a global health, wellbeing and science R&D brief.** **338**
Question . . . 338
Background . . . 338
Sub-Question 1: **Cost-Effectiveness Logic and What Comparable Actually Claims** . . . 339
Sub-Question 2: **From Grant to Expected Deaths: the AMF / DRC Case** . . . 341
Sub-Question 3: **Market-Failure Justifications: Which Failure, and Is It Binding?** . . . 343
Sub-Question 4: **Measurement Caveats: Attribution, Displacement and Counterfactuality Are Three Different Errors** . 345

**Brief on farm animal welfare and Effective Giving & Careers.** **349**
Question . . . 349

Background . . . 349
Sub-Question 1: Farm Animal Welfare: Chain of Leverage, Counting Unit, and the Persistence Conditional . . . 349
Sub-Question 2: Effective Giving: Which Return on Investment, and Why the Same Word Means Two Different Things . . . 353
Sub-Question 3: Effective Careers: The RFP as Funder Instrument, Outputs Versus Outcomes, and Selection on Measurability . . . 356
Sub-Question 4: Cross-Cause Comparison: Counterfactuality, Persistence, Credit Allocation, and the Missing Common Currency . . . 360

**Analyze abundance, growth and catch-up growth after AI. 364**
Question . . . 364
Background . . . 364
Sub-Question 1: **What is growth as a CG cause area, and why are "abundance" and "catch-up growth" different objects?** 365
Sub-Question 2: **Which assumptions in the catch-up argument are load-bearing, and is the reported arithmetic internally consistent?** . . . 366
Sub-Question 3: **How much of the headline survives a change in D, and what happens to the recommendation when D is 20 rather than 10?** . . . 369
Sub-Question 4: **What are the main objections, how would a grantmaker adjudicate them, and why does this matter to a GCR or generalist evaluator?** . . . 372

**Reason about cross-cause prioritisation and worldview diversification. 375**
Question . . . 375
Background . . . 375
Sub-Question 1: **Marginalist Funding and the Cross-Cause Comparison Problem** . . . 376
Sub-Question 2: **Worldview Diversification versus Focused Cause Prioritisation** . . . 379
Sub-Question 3: **General-Manager Ownership and the Tension with Marginalist Funding** . . . 381
Sub-Question 4: **Worked Exercise: Allocate a Marginal 10 Million Dollars Across Three Cause Areas** . . . 383

**Understand policy process, political feasibility and implementation. 387**
Question . . . 387
Background . . . 387
Sub-Question 1: The Anatomy of Political Infeasibility: Four Rejection Mechanisms, Four Cost Bearers . . . 388
Sub-Question 2: The Staged Model as a Multiplicative Chain: Locating the Binding Gate . . . 391

Sub-Question 3: Windows, Entrepreneurs, Venues and Frames: The Four Levers That Move the Gates . . . 395
Sub-Question 4: Venues Are the Real Decision Variable: The Three Governance Pathways and the Post-Enactment Chain . . . 399
Sub-Question 5: Worked Judgment Exercise: Diagnosing the Four-Pillar Biosecurity Plan by Venue . . . 403

**Build a crux table of disagreements within AI safety and EA philanthropy. 408**
Question . . . 408
Background . . . 408
Sub-Question 1: Method: the crux test and the attribution ladder . . . 408
Sub-Question 2: The crux table . . . 410
Sub-Question 3: Cross-cutting load-bearing assumptions and the evidence that would move each side . . . 428
Sub-Question 4: Reading the table as an evaluator . . . 430

**Map institutional differences among funders and organizations. 432**
Question . . . 432
Background . . . 432
Sub-Question 1: **The comparison schema: funders as decision architectures, not points on a scale** . . . 432
Sub-Question 2: **Scale and speed: deployment architecture and the latency-size tradeoff** . . . 438
Sub-Question 3: **Risk tolerance and transparency: two axes that are not the same** . . . 440
Sub-Question 4: **Grantee access and dealflow: the funnel, not the list** . . . 441
Sub-Question 5: **Institutional versus individual versus third-party: weighting the sources** . . . 443
Sub-Question 6: **Worked exercise: matching two hypothetical opportunities to funders** . . . 446

**Create an uncertainty register for major claims. 449**
Question . . . 449
Background . . . 449
Sub-Question 1: **Subquestion 1: Register architecture, confidence vocabularies and tagging discipline** . . . 450
Sub-Question 2: **Subquestion 2: Timelines and takeoff-speed entries** . . . 454
Sub-Question 3: **Subquestion 3: AI-risk entries and the openness, China and politics disagreement cluster** . . . 459
Sub-Question 4: **Subquestion 4: Biorisk entries and the opposing-evidence crux** . . . 467
Sub-Question 5: **Subquestion 5: Intervention effectiveness and ROI entries, and the comparability problem** . . . 477

Sub-Question 6: **Subquestion 6: Catch-up growth entries and the fragility of the central calibration** . . . 485
Sub-Question 7: **Subquestion 7: Crux atlas, influence ranking and calibrated interview usage** . . . 492

**Catalogue counterexamples and failure modes in grantmaking.** **496**
Question . . . 496
Background . . . 496
Sub-Question 1: **Locus of Failure: Decomposing 'It Failed' into an Attributable Claim** . . . 497
Sub-Question 2: **Counterexamples: What Each Case Licenses and What It Does Not** . . . 500
Sub-Question 3: **Mitigations and a Worked Diagnosis: Turning the Catalogue into Evaluator Behaviour** . . . 504

**Build a likely-question bank and answer outlines.** **509**
Question . . . 509
Background . . . 509
Sub-Question 1: **Requirement-to-Question-Category Map** . . . 509
Sub-Question 2: **Motivation, Role Fit, Strategy and Scale** . . . 512
Sub-Question 3: **Grant Evaluation, Due Diligence and Judging People** . . . 519
Sub-Question 4: **Quantitative and Probabilistic Reasoning** . . . 526
Sub-Question 5: **AI Safety, Security, Governance and Biosecurity** 533
Sub-Question 6: **Institutional Differences, Disagreements and Cruxes** . . . 546
Sub-Question 7: **Communication, Calibration and Reasoning Transparency** . . . 553
Sub-Question 8: **Alternative Defensible Answers for Selected Questions** . . . 560
Sub-Question 9: **Rehearsal Protocol and Stress Tests** . . . 568

**Complete worked grant-evaluation drills with alternative defensible answers.** **570**
Question . . . 570
Background . . . 570
Sub-Question 1: **The memo schema, the rubric, and the three invented opportunities** . . . 572
Sub-Question 2: **Memo 1 - large AI safety scale-up: the dollar amount is not the binding question** . . . 578
Sub-Question 3: **Memo 2 - sub-$10k rapid grant: the evaluator's own cost structure is part of the answer** . . . 585
Sub-Question 4: **Memo 3 - policy and field-building: pay for the cheapest relative improvement, not the most persuasive stage** . . . 591

Sub-Question 5: **Sensitivity: the same three memos under short versus long timelines and high versus low risk tolerance** . 598
Sub-Question 6: **Alternative defensible answers, reversal conditions, and final calibration** . . . . . 603

**Draft calibrated answers for comparing opposing arguments and expected value.** **608**
Question . . . . . 608
Background . . . . . 608
Sub-Question 1: **The answer skeleton: one question, one crux, three numbers, one falsifier** . . . . . 609
Sub-Question 2: **Scripts for competing expert claims, with locus preserved** . . . . . 612
Sub-Question 3: **Scripts for expected value: unit discipline before arithmetic** . . . . . 615
Sub-Question 4: **Scripts for decisive considerations and proportionate due diligence** . . . . . 619
Sub-Question 5: **Adversarial calibration: reversing each script and stating the limits** . . . . . 621

**Practice concise communication and reasoning transparency.** **625**
Question . . . . . 625
Background . . . . . 625
Sub-Question 1: **Build the Communication Contract: D01 plus L02 as a Checklist** . . . . . 626
Sub-Question 2: **Three 60-120 Second Spoken Answers** . . . . . 628
Sub-Question 3: **Two Short Written Memos** . . . . . 630
Sub-Question 4: **Rehearsal, Feedback and AI-Writing Guardrails** 632

**Rehearse institutional-disagreement and uncertainty questions.** **634**
Question . . . . . 634
Background . . . . . 634
Sub-Question 1: Drill 0 — the rehearsal form and the attribution discipline . . . . . 635
Sub-Question 2: Drill A — are short timelines the default path? . . . 637
Sub-Question 3: Drill B — centralised staffed grantmaking versus dispersed regrantors . . . . . 640
Sub-Question 4: Drill C — two funded papers that point in opposite directions . . . . . 644
Sub-Question 5: Drill D — does the funder accelerate capability, and what is its theory of change? . . . . . 646
Sub-Question 6: Drill E — what does the funder think? separating institution from individual . . . . . 649
Sub-Question 7: Drill F — how confident are you, and what would change your mind? . . . . . 651

Sub-Question 8: Capstone — the ninety-minute rehearsal circuit and its limits . . . 653

**Write the final one-page positioning note and source index. 656**
Question . . . 656
Background . . . 656
Sub-Question 1: **Draft the positioning note and the roughly 200-word answer** . . . 657
Sub-Question 2: **Build the concise source index and the four-way separation map** . . . 660
Sub-Question 3: **Run the fabrication and status check, then freeze the one-pager** . . . 664

**Run a mock interview and feedback loop. 668**
Question . . . 668
Background . . . 668
Sub-Question 1: **Subquestion 1: Build the instrument and verify coverage** . . . 669
Sub-Question 2: **Subquestion 2: Run the first pass and produce the transcript** . . . 671
Sub-Question 3: **Subquestion 3: Score the first pass with gates before dimensions** . . . 674
Sub-Question 4: **Subquestion 4: Triage, revise, and run the second pass on fresh prompts** . . . 678

# Decode the attributed role brief and extract its explicit and implicit evaluation criteria.

## Question

Given an attributed role brief that is explicitly not a reproduction of the complete advertisement, what evaluation criteria can be responsibly extracted, how do those criteria map to likely interview evidence, and where must uncertainty be preserved?

This step decodes the brief for preparation purposes. It does not claim a verified rubric or a current vacancy.

## Background

Source status: the file evaluator-job-ad-input-2026-09-16.txt is an attributed role brief. It states it is not a reproduction of the complete advertisement. Employer: Coefficient Giving. Advertisement: Grantmakers & Senior Generalists, Global Catastrophic Risks. URL recorded. Retrieved 2026-09-16. The browser-visible deadline of May 24, 2026 is historical. It does not establish a current vacancy or an application by the learner.

Content scope: grantmaking in AI governance, catastrophic-risk capacity building, short-timeline projects and biosecurity. Grantmakers investigate funding opportunities, develop neglected projects, recruit founders, launch initiatives, research program strategy and historical cost-effectiveness, and maintain grantee and stakeholder relationships.

Evaluation signals in the brief: comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and judging people and organizations. Communication should be concise, calibrated and transparent about reasoning.

Relevant backgrounds: policy, frontier-AI organizations, technical fields, or experience within AI safety, biosecurity and associated research or nonprofit communities.

Learning intent: do not invent an application, credentials, work history or experience. Preserve uncertainty and disagreement. Connect recommendations to sources. Include worked judgment exercises with alternative defensible answers.

This document treats the brief as orientation. It separates source-supported claims from inferences and proposed exercises.

## Sub-Question 1: Source Status and Epistemic License

What does the provenance of the attributed role brief permit us to infer about evaluation criteria, and what does it forbid?

The answer determines whether later mapping is a reconstruction of stated requirements or an overread of a partial artifact.

**Step 1**

The brief mixes three kinds of statement: provenance facts, role-content claims, and learning instructions. Treat each with different evidential weight.

Table 1: Provenance and content are separate; do not collapse them.

| **Statement type** | **Source location** | **What it supports** | **What it does not support** |
|---|---|---|---|
| Provenance: employer, title, URL, retrieval date, historical deadline | PRIMARY SOURCE | The artifact is an attributed brief referencing a CG advertisement. | It does not prove the advertisement is currently open or that the learner applied. |
| Role-content: grantmaking domains and activities | ATTRIBUTED ROLE SUMMARY | Orientation to the role's scope. | It is not a complete duty statement or scoring rubric. |
| Evaluation-content: comparing arguments, quantitative expected value, decisive considerations, due diligence, judging people | ATTRIBUTED ROLE SUMMARY | Likely evaluation dimensions for interview preparation. | The exact weighting, format, and sample questions are not supplied. |
| Background-content: policy, frontier AI, technical fields, AI safety and biosecurity communities | ATTRIBUTED ROLE SUMMARY | Possible background signals. | It does not define minimum qualifications or required credentials. |
| Learning intent: no invented application or credentials; preserve disagreement | LEARNING INTENT | Constraints on how to prepare. | It does not remove the need to infer from other CG sources. |

**Step 2**

The brief is best treated as a high-signal orientation document. It is source-supported that these themes appear in the attributed summary. It is an inference that they constitute the interview rubric. It would be speculation to assign weights or predict exact questions.

- Source-supported: the brief says evaluation requires comparing opposing arguments.
- Source-supported: the brief says evaluation requires quantitative and probabilistic reasoning about expected value.
- Inference: these are likely interview criteria because they are listed in the role summary.
- Speculation: a specific scoring formula or weighting.
- Speculation: a particular interview format.

**Step 3**

Prepare to demonstrate the stated competencies through real, non-invented evidence. Where the learner lacks direct experience, prepare analytical reasoning and explicit calibration rather than fabricated history. Flag uncertainty about the rubric where relevant.

## Sub-Question 2: Explicit Criteria and Likely Interview Evidence

What are the explicit requirements in the attributed brief, and what interview evidence would plausibly bear on each?

The mapping is a hypothesis about evidence, not a claim about the actual interview.

**Step 1**

Table 2: Explicit requirements mapped to plausible evidence and gaps.

| **Explicit requirement** | **Source phrase or summary** | **Likely interview evidence** | **Failure mode to avoid** | **Gap or uncertainty** |
|---|---|---|---|---|
| Ownership of ambiguous work | Role calls for ownership of ambiguous work | A real example where the learner defined a problem, chose a path under uncertainty, and delivered; analytical case where the learner structures an ambiguous prompt. | Waiting for instructions; overclaiming certainty; inventing experience. | The brief does not define the scope or level of ambiguity expected. |
| Ambition and speed | Role calls for ambition and speed | Evidence of rapid prioritization, fast but proportionate decisions, and willingness to make big bets; mock grant triage under time pressure. | Recklessness; confusing speed with haste; ignoring due diligence. | No stated threshold for speed or ambition. |

| **Explicit requirement** | **Source phrase or summary** | **Likely interview evidence** | **Failure mode to avoid** | **Gap or uncertainty** |
|---|---|---|---|---|
| Interpersonal adaptability | Role calls for interpersonal adaptability | Examples of adjusting to stakeholders, grantees, founders, technical experts, and policymakers; handling disagreement. | Rigidity; over-accommodation; inability to say no. | The brief does not specify stakeholder mix or conflict scenarios. |
| Familiarity with catastrophic-risk communities | Role calls for familiarity with catastrophic-risk communities | Demonstrated knowledge of AI safety, biosecurity, governance, and field norms; accurate use of community concepts. | Name-dropping; treating the field as monolithic; merging organizations into one voice. | The brief does not say whether deep technical or policy expertise is required. |
| Comparing opposing arguments | Evaluation requires comparing opposing arguments | Worked example where the learner steelmans two sides, states cruxes, and updates. | Strawmanning; false balance; refusing to conclude. | No supplied example of a target comparison. |
| Quantitative and probabilistic expected value | Reasoning quantitatively and probabilistically about expected value | Back-of-the-envelope calculation with explicit assumptions, ranges, and sensitivity; probability estimates with calibration caveats. | False precision; hiding assumptions; ignoring uncertainty. | The brief does not state required mathematical depth. |

| **Explicit requirement** | **Source phrase or summary** | **Likely interview evidence** | **Failure mode to avoid** | **Gap or uncertainty** |
|---|---|---|---|---|
| Identifying decisive considerations | Identifying decisive considerations | A case where the learner distinguishes decisive from merely relevant factors and explains why. | Listing every consideration; missing the crux; confusing salience with importance. | No stated decision framework. |
| Proportionate due diligence | Choosing proportionate due diligence | A plan that scales investigation to grant size, risk, and reversibility; stopping rules. | Over-investigating small grants; under-investigating large or risky grants. | The brief does not supply CG's actual thresholds. |
| Judging people and organizations | Judging people and organizations | Reference-check reasoning, track-record interpretation, founder evaluation, organizational capacity assessment. | Overconfidence from thin evidence; halo effects; ignoring base rates. | The brief does not state which judgment tools CG uses. |
| Concise, calibrated, transparent communication | Communication should be concise, calibrated and transparent about reasoning | A short memo or verbal answer with key takeaways, confidence levels, and linked reasoning. | Verbosity; hidden uncertainty; overclaiming. | The brief does not give a desired length or format. |

**Step 2**

Table 3: Clusters are inferred groupings for preparation, not stated categories in the brief.

| **Cluster** | **Requirements in cluster** | **What a strong answer demonstrates** | **What a weak answer reveals** |
|---|---|---|---|
| Agency and ownership | Ownership of ambiguous work; ambition and speed; launching initiatives; recruiting founders | Proactive problem definition, bias to action, willingness to make and defend judgment calls. | Passivity, waiting for a rubric, reluctance to choose. |
| Analytical evaluation | Comparing opposing arguments; quantitative and probabilistic expected value; decisive considerations; proportionate due diligence; judging people and organizations | Structured reasoning, explicit assumptions, calibrated confidence, cost-effectiveness thinking. | Unstructured opinion, false precision, missing cruxes. |
| Domain fluency | Familiarity with catastrophic-risk communities; AI governance; biosecurity; capacity building; short-timeline projects | Accurate distinctions among actors, agendas, and uncertainties. | Sloganeering, merging distinct positions, outdated claims. |
| Communication and stakeholder craft | Concise, calibrated, transparent communication; maintaining relationships | Short, legible reasoning; appropriate caveats; stakeholder sensitivity. | Long-windedness, opacity, or unnecessary hedging. |

**Step 3**

- The brief is attributed, not a complete advertisement, so the explicit list may omit requirements.

- The brief does not provide weights, scoring anchors, or sample interview questions.
- The brief does not specify whether the role is more evaluator or more grantmaker at entry.
- The brief does not state how much technical AI or biosecurity knowledge is required versus learnable.
- The historical deadline means current hiring signals must come from elsewhere, such as P04, and even those are historical.

## Sub-Question 3: Implicit Criteria from Institutional Context

What evaluation criteria are not stated in the brief but are implied by Coefficient Giving's public texts and by the structure of the role?

These inferences must be labeled as inferences and tied to specific sources.

**Step 1**

Table 4: Implicit criteria inferred from CG texts; these are not directly stated in the attributed brief.

| **Implicit criterion** | **Source anchor** | **Why it is implied** | **Confidence** | **What would raise confidence** |
|---|---|---|---|---|
| Reasoning transparency | D01 Reasoning Transparency | CG publicly values analyses that show key takeaways, importance, confidence, and support. | High for CG culture; medium for this role | A role-specific writing sample or interview prompt. |
| Back-of-the-envelope calculation and cost-effectiveness | D02 How We Use AI to Speed Up Our Work; P01; P06 | CG grantmakers use BOTECs and cost-effectiveness framing; P01 discusses scaling and prioritization. | High for grantmaking; medium for evaluator entry | A stated case-study format in the interview. |

| **Implicit criterion** | **Source anchor** | **Why it is implied** | **Confidence** | **What would raise confidence** |
|---|---|---|---|---|
| Marginalist versus infra-marginal judgment | P06 2025 Letter from the CEO | CG describes a shift from equalizing marginal returns toward proactive, infra-marginal bets while retaining marginalist discipline. | High for institutional strategy; medium for role criteria | A prompt about portfolio allocation or founder recruitment. |
| Proactive grantmaking and gap-filling | P01; P04; P06; S09 Project Tailwind | CG emphasizes proactive identification, seeding organizations, and recruiting founders. | High | A behavioral question about creating opportunities. |
| Capacity constraint and speed | P01; P04; P06 | CG repeatedly says staff time is the bottleneck and that faster decisions matter. | High | A triage or prioritization exercise. |

| Implicit criterion | Source anchor | Why it is implied | Confidence | What would raise confidence |
|---|---|---|---|---|
| Donor and advisory orientation | S07; P10; S08 | CG is scaling partnerships and expects growth in philanthropic capital; advisory work is a leadership priority. | Medium for this role; high for institution | A question about recommending grants to other funders. |
| Judging people and organizations | P04; S05; S03 | Grantmakers describe reference checks, founder recruitment, and incubation pipelines. | High | A reference-check or founder-evaluation scenario. |
| Field norms and community familiarity | P04; P05; P10; P11 | Grantmakers work with grantees, fellowships, founders, and community institutions. | High | A question about navigating community disagreement. |

**Step 2**

Table 5: Inferences from wording and omissions are weaker than source-supported claims.

| Wording or omission | Implicit signal | Alternative reading | Caution |
|---|---|---|---|
| Ownership of ambiguous work | Comfort with unstructured problems and personal accountability. | The role may provide more structure than the phrase suggests. | Do not assume no supervision or no process. |

| Wording or omission | Implicit signal | Alternative reading | Caution |
|---|---|---|---|
| Ambition and speed | Preference for action and scale over caution. | CG also emphasizes proportionate due diligence and risk awareness. | Speed is not an excuse for weak reasoning. |
| Interpersonal adaptability | Stakeholder-heavy work with founders, grantees, experts, and donors. | May also require internal navigation and saying no. | Do not reduce to likeability. |
| Catastrophic-risk familiarity | Community fluency and network matter. | The brief also lists policy, frontier labs, and technical fields. | Do not treat one community as the whole field. |
| Judging people and organizations | People judgment is a core evaluation skill. | It may be exercised through structured evidence and references rather than intuition. | Avoid overconfidence from interviews alone. |
| Omission of exact rubric | The brief is orientation, not a scoring sheet. | The real process may weight criteria differently. | Preserve uncertainty. |

**Step 3**

- Source-supported: the brief lists the explicit requirements in the ATTRIBUTED ROLE SUMMARY.
- Source-supported: CG's public texts describe reasoning transparency, BOTECs, proactive grantmaking, and scale.
- Inferred: these institutional themes are likely interview criteria for this role.
- Inferred: implicit criteria include marginalist discipline and capacity-constrained speed.
- Speculative: specific weights, sample questions, and scoring anchors.

## Sub-Question 4: Gaps, Alternative Readings, and Preparation Conversion

What remains unknown, what alternative interpretations of the brief are defensible, and how should these criteria be converted into preparation without inventing credentials?

This section turns decoding into action while keeping uncertainty visible.

### Step 1

Table 6: Gaps are not failures of preparation; they are explicit uncertainties to manage.

| **Gap or uncertainty** | **Why it matters** | **What would resolve it** | **How to handle it now** |
|---|---|---|---|
| No complete advertisement | The brief may omit requirements or context. | The original advertisement or a current job description. | Treat the brief as orientation and seek current signals from CG hiring pages. |
| Historical deadline | It does not establish a current vacancy. | A current posting or recruiter confirmation. | Do not assume an open role; prepare for the role type. |
| No rubric or weights | Cannot know which competencies are decisive. | Interview format, scoring guide, or feedback. | Prepare across clusters and state assumptions. |
| No sample tasks | Cannot know the exact analytical format. | Past interview accounts or practice prompts. | Rehearse BOTECs, written memos, and verbal reasoning. |
| No definition of evaluator versus grantmaker split | The role may emphasize origination, evaluation, or both. | Team structure and role level. | Prepare both evaluative and proactive grantmaking examples. |
| No stated technical depth | Could be policy-generalist or technical specialist. | Role level and team placement. | Calibrate depth to the source material and avoid false expertise. |

| Gap or uncertainty | Why it matters | What would resolve it | How to handle it now |
|---|---|---|---|
| No information on current funds | Fund boundaries and scale may have shifted. | Current CG fund pages and recent posts. | Cite source dates and avoid outdated figures as current. |

**Step 2**

- Reading A: The brief is a literal competency list. Under this reading, prepare direct evidence for each stated requirement.
- Reading B: The brief is a thematic orientation. Under this reading, prepare clusters and reasoning rather than one-to-one evidence.
- Reading C: The role is a generalist grantmaker with evaluator duties. Under this reading, emphasize ownership, speed, stakeholder craft, and judgment.
- Reading D: The role is a specialist evaluator within a domain. Under this reading, emphasize domain fluency and analytical depth.
- Reading E: The role is a donor-facing advisory role. Under this reading, emphasize communication, calibration, and funder relations.

These readings are not mutually exclusive. The strongest preparation covers the overlap while making assumptions explicit.

**Step 3**

Table 7: Preparation uses real evidence or analytical rehearsal, never invented history.

| Competency | Acceptable evidence type | Not acceptable | Preparation action |
|---|---|---|---|
| Ownership of ambiguous work | Real past examples of defining a problem and delivering under uncertainty; otherwise analytical case responses. | Fabricated job history or titles. | Write three STAR-style skeletons from real experience; if none, write analytical case notes. |

| **Competency** | **Acceptable evidence type** | **Not acceptable** | **Preparation action** |
|---|---|---|---|
| Ambition and speed | Real examples of rapid prioritization or initiative; otherwise mock triage exercises. | Claiming experience that does not exist. | Practice time-boxed prioritization with explicit tradeoffs. |
| Interpersonal adaptability | Real stakeholder examples; otherwise role-play and conflict scenarios. | Invented stakeholder relationships. | Rehearse difficult conversations and saying no with reasons. |
| Catastrophic-risk familiarity | Source-anchored notes on AI safety, biosecurity, governance, and field disagreements. | Name-dropping without understanding. | Build a one-page map of actors, agendas, and uncertainties. |
| Comparing opposing arguments | Worked examples using supplied sources. | Strawmanning a position. | Write a two-column steelman and identify cruxes. |
| Quantitative and probabilistic expected value | BOTEC practice with explicit assumptions and ranges. | False precision. | Do three BOTECs: grant cost-effectiveness, founder bet, portfolio allocation. |
| Decisive considerations | Case analyses that isolate the crux. | Listing every factor. | Practice writing 'the decision turns on X because Y'. |
| Proportionate due diligence | Due diligence plans scaled to stakes. | One-size-fits-all diligence. | Create a triage rubric for small, medium, and large grants. |
| Judging people and organizations | Structured reference-check reasoning and track-record analysis. | Unsupported intuition. | Practice evidence-to-judgment chains with base rates. |

| Competency | Acceptable evidence type | Not acceptable | Preparation action |
|---|---|---|---|
| Concise, calibrated communication | Short memos and verbal answers with confidence labels. | Verbose, hidden uncertainty. | Write 150-word answers with key takeaway, confidence, and what would change your mind. |

Worked judgment exercise: use the CG AI Governance RFP criteria (D04) to decide whether to fund a new AI governance research project. Give two defensible answers and the evidence that would break the tie.

Table 8: Both answers are defensible; the tie-breaker is often decision-relevant evidence, not ideology.

| Decision element | Answer A: fund | Answer B: do not fund yet | What would break the tie |
|---|---|---|---|
| Theory of change | The project addresses a neglected governance bottleneck and has a plausible path to policy influence. | The path to influence is speculative and lacks a specific decision-maker audience. | Evidence of a named audience and a concrete policy window. |
| Track record | The applicant has relevant publications or prior project delivery. | The applicant lacks a track record in this exact domain. | Reference checks with domain experts. |
| Strategic judgment | The proposal shows awareness of failure modes and alternative approaches. | The proposal underweights political feasibility and implementation. | A revised proposal that addresses feasibility. |
| Project risks | Risks are identified and proportionate to the grant size. | Downside risks are understated or unaddressed. | A risk register with mitigation and stop conditions. |

| **Decision element** | **Answer A: fund** | **Answer B: do not fund yet** | **What would break the tie** |
|---|---|---|---|
| Cost-effectiveness | The budget is proportionate to the expected decision-relevant output. | The budget is high relative to the uncertainty of impact. | A smaller scoped pilot with clear milestones. |
| Scale | A larger grant is justified by evidence of early success. | Start with a smaller grant until evidence accrues. | A stage-gated funding plan. |

1. State the question and your interpretation of the decision.
2. Give the key takeaway in one sentence.
3. State your confidence and the main assumptions.
4. Identify the decisive consideration or crux.
5. Give the strongest opposing argument and what would change your mind.
6. State the proportionate next step, including diligence or a small bet.
7. Flag what you do not know and how you would learn it.

This template operationalizes concise, calibrated, reasoning-transparent communication without inventing credentials.

# Map Coefficient Giving's structure, funds, giving scale and strategic posture.

## Question

Map Coefficient Giving's structure, funds, giving scale, and strategic posture. The deliverable is a one-page fund and scale map that distinguishes the GCR AI safety and security portfolio, the biosecurity portfolio, Effective Giving & Careers, Global Health and Wellbeing, Farm Animal Welfare, Abundance & Growth, Global Growth, and other funds, with approximate giving figures and implications for an evaluator or grantmaker.

## Background

This map is a source-anchored reconstruction of Coefficient Giving (CG) as described in the supplied corpus. It is not a current org chart, and it deliberately preserves the difference between the attributed role brief, CG's own public claims, individual staff views, and third-party or adjacent actors' claims.

Scale claims split across two units of analysis. P01 reports the AI safety, security, and fieldbuilding portfolio at $168m in 2024, $351m in 2025, and on track to commit over $1b in 2026, with $352m committed in August alone; biosecurity grew from $30m in 2024 and $65m in 2025 to more than $150m so far this year. P06 reports that CG directed over $1b in total in 2025, increased total giving by over 50%, launched two new funds, rebranded, and more than doubled funding from non-Good Ventures donors to over $200m. These are different denominators.

Fund-level anchors include P05's Effective Giving & Careers fund with more than 60 grants and over $30m since 2022 and nearly 3x funding available in 2026; P07's staff suggestions across global health and wellbeing, science and global health R&D, farm animal welfare, global aid policy, abundance and growth, navigating transformative AI, and effective giving and careers; and P06's mention of the Abundance & Growth Fund, Global Growth Fund, biosecurity, AI, global health and wellbeing, and farm animal welfare. P10 and P11 describe adjacent external vehicles, the EA Funds Transformative AI Fund and Manifund regranting, which matter for co-funding and gap analysis.

Critical caveats: 'committed' grants are formally recommended grants and need not match public grants databases; conditionally approved grants are counted separately; figures are source claims that may be outdated; the institution formerly published as Open Philanthropy; P07 individual donor suggestions are not institutional views; and the role brief and P04 hiring deadlines are historical. The map therefore functions as a reasoning aid for an interview, not as a live financial statement.

## Sub-Question 1: Fund Architecture and Giving Scale

Build the one-page map: identify each fund or portfolio, its domain, the approximate giving anchors, the source anchors, and the evaluator or grantmaker implication.

Keep total CG giving separate from portfolio-level giving, and keep source-supported figures separate from your inference about where new money is going.

### Step 1

Table 9: Table 1. Source-anchored fund and scale map. Figures are source claims, not audited totals.

| **Portfolio or fund** | **Primary domain** | **Approximate giving anchors** | **Source anchors** | **Evaluator or grant-maker implication** |
|---|---|---|---|---|
| **AI safety and security** | Technical AI safety, security, fieldbuilding, AI governance and policy, navigating transformative AI | $168m in 2024; $351m in 2025; on track over $1b in 2026; $352m committed in August alone | P01; P06; P08; D04 | Largest and fastest-growing portfolio; an evaluator must know technical safety, security, governance, policy, fieldbuilding, and proactive founder seeding |

| Portfolio or fund | Primary domain | Approximate giving anchors | Source anchors | Evaluator or grant-maker implication |
|---|---|---|---|---|
| **Biosecurity and pandemic prepared-ness** | Biosecurity, pandemic prepared-ness, AIxBio intersection | $30m in 2024; $65m in 2025; more than $150m so far in 2026 | P01; P04; P06 | Rapid scale-up with distinct grantmaker workflow and RFP capacity; evaluator should separate biosecurity from AI safety while tracking the intersection |
| **Effective Giving & Careers** | Effective giving, talent, incubation, research | More than 60 grants and over $30m since 2022; $13,075,301 to 28 organi-zations in first effective careers RFP; nearly 3x funding available in 2026 | P02; P05; P07 | Different ROI and counterfac-tuality logic; evaluator should understand cost-effective giving, career transitions, incubation, and M&E |

| Portfolio or fund | Primary domain | Approximate giving anchors | Source anchors | Evaluator or grant-maker implication |
|---|---|---|---|---|
| **Global Health and Wellbeing** | Evidence-backed global health and wellbeing | P06 says grants to evidence-backed global health interventions have saved over 100,000 lives; S01 reports $212m to Against Malaria Foundation plus $64m from other GiveWell donors; P06 reports over $1b total in 2025 | P06; P07; S01; S04 | Not the same bar or evidence model as GCR; evaluator should know GiveWell-style evidence and cost-effectiveness |
| **Farm Animal Welfare** | Farm animal welfare | P06 says work has improved lives of over 3 billion farm animals and projects another 10 billion over 10 years; P07 recommends EA Animal Welfare Fund | P06; P07 | Corporate campaigns and supply-chain commit-ments; different theory of change and measure-ment |

| Portfolio or fund | Primary domain | Approximate giving anchors | Source anchors | Evaluator or grant-maker implication |
|---|---|---|---|---|
| **Abundance & Growth** | Housing, growth, abundance | P06 says launched as new fund with Matt Clancy and others; P07 recommends Center for Building in North America | P06; P07 | Domestic policy and growth portfolio; evaluator should not import GCR reasoning wholesale |
| **Global Growth** | Economic growth | P06 says Justin Sandefur joined from Center for Global Development to lead new Global Growth Fund; P03 discusses catch-up growth after AI | P06; P03 | New fund connecting growth, AI, and global welfare; a source of cross-portfolio reasoning |

| Portfolio or fund | Primary domain | Approximate giving anchors | Source anchors | Evaluator or grant-maker implication |
|---|---|---|---|---|
| **Other, adjacent, and external vehicles** | Science and global health R&D, global aid policy, individual donor suggestions, EA Funds Transformative AI Fund, Manifund regranting, AI governance RFP referrals | P07 lists science and global health R&D, global aid policy, and individual suggestions; P10 describes about $1m seed for Transformative AI Fund; P11 describes $2m 2025 Manifund regranting pool and target $5m; D04 says more than 20 other philanthropists may fund recommendations and their giving may exceed $50m in 2024 | P07; P10; P11; D04 | CG is a hub in a wider funding ecosystem; evaluator should know where CG stops and others start |

**Step 2**

- Non-trivial read 1: total CG giving is not the AI portfolio. P06's over $1b in 2025 is institution-wide; P01's $351m in 2025 is the AI safety, security, and fieldbuilding portfolio. Do not use one as the other.
- Non-trivial read 2: 'committed' is a recommendation construct. P01 footnote 2 says committed grants are formally recommended and approximate money reasonably sure to reach grantees; conditionally approved grants

are counted separately; neither exactly matches public grants databases.

- Non-trivial read 3: fund boundaries overlap. AI safety includes technical safety, security, governance, policy, and fieldbuilding; biosecurity includes AIxBio; Effective Giving & Careers spans cause areas; P07 areas cut across funds.
- Non-trivial read 4: growth is uneven and strategic. P01 says AI and biosecurity are scaling fastest; P05 says Effective Giving & Careers has nearly 3x funding in 2026; P06 says nearly every fund has a higher budget. The evaluator should infer where marginal grantmaker time is being pushed.
- Non-trivial read 5: external funds are part of the map. P10's Transformative AI Fund and P11's Manifund regranting occupy early-stage and small-grant niches; D04's referral network means a CG recommendation may be funded by Good Ventures or more than 20 other philanthropists.

**Step 3**

Implication for a one-page map: use three columns of interpretation. First, scale: total CG versus portfolio versus single grant. Second, mechanism: proactive seeding, RFP, active grantmaking, renewal, or referral. Third, fit: which fund's evidence standard and theory of change applies. This prevents the common interview error of treating all CG money as governed by one GCR cost-effectiveness bar.

## Sub-Question 2: Strategic Posture and Evaluator Implications

Map the posture that produces the fund structure: urgency, speed, proactive ownership, marginalist discipline, donor scale, agility, and staffing constraint.

Then convert each posture into an evaluator or grantmaker implication, and mark what remains uncertain.

**Step 1**

Table 10: Table 2. Strategic posture and its evaluator implications.

| Posture | Evidence | Evaluator or grantmaker implication |
|---|---|---|
| **Urgency and short timelines** | P01 says transformative AI may arrive soon, cites shortened timelines, and says short-timeline predictors look increasingly correct; P06 says short AGI timelines have been more accurate than the CEO's own, and Claire Zabel leads a short-timelines team | Grantmaker must reason under time pressure and hedge across timelines |
| **Speed and delegation** | P01 reports shortened grant investigations, approval from one additional person, and a $10m grant approved 12 days after first conversation; P06 says process changes let teams move faster | Evaluator must be able to make or support fast, proportionate due diligence |
| **Proactive seeding** | P01 describes Project Tailwind and seeding new organizations; S09 lists pre-seed $200k-$2m, seed $2m-$20m, scale $20m-$200m+; D04 says AI governance RFP shifted toward proactive gap identification | Grantmaker is not only reactive; must source founders and create organizations |

| Posture | Evidence | Evaluator or grantmaker implication |
|---|---|---|
| **Marginalist versus general manager** | P06 describes historical marginalist mode of equalizing marginal returns, then a shift toward ambitious ownership and inframarginal bets, while warning against the strategic funder trap | Evaluator must balance cost-effectiveness discipline with ownership and local grantee knowledge |
| **Donor and advisory scale** | P06 says non-Good Ventures funding more than doubled to over $200m and aims to prepare for larger growth; D04 says more than 20 philanthropists and referrals; S07 describes a Managing Director, Philanthropic Advisory role | Grantmaker may need to write for multiple funders and preserve optionality |
| **Agility and culture** | P06 cites Netflix-style minimal bureaucracy, pushing autonomy down, shorter processes, and AI enablement; D02 describes AI tools for grant writeups and BOTECs | Evaluator should be tool-fluent and reason transparently, not hide behind process |
| **Staffing constraint** | P01 says CG remains capacity-constrained and staff time is the main limit; P06 says aims to bring on about 70 new staff; D04 says RFP volume overwhelmed capacity and raised the bar | Grantmaker must prioritize opportunities and decline good proposals |

**Step 2**

- Read the role brief as a fit test: ownership of ambiguous work, ambition and speed, interpersonal adaptability, familiarity with catastrophic-risk communities, comparing opposing arguments, quantitative and probabilistic expected value, decisive considerations, proportionate due diligence, and judging people and organizations.
- Translate fund structure into questions you can ask: Which portfolio would this grant sit in? What is that portfolio's bar? Who else is funding this niche? Is this a proactive seed, an RFP response, a renewal, or a referral?
- Use calibration language: state probabilities or confidence intervals where the source does, and distinguish source-supported claims from your inference and from proposed exercises.
- Watch the denominator: a grant can be large relative to a small field but small relative to CG's total giving; a fund can grow fast from a low base; a recommendation can be funded by Good Ventures or by other philanthropists.
- Preserve disagreement: P01, P06, P08, D04, P10, P11, and individual staff do not speak with one voice; the role rewards reasoning transparency about those differences.

**Step 3**

- The role brief and P04 are historical; they do not establish a current vacancy or an application.
- Giving figures are source claims that may be outdated, and 'committed' differs from conditionally approved and published grants.
- Fund names and boundaries have shifted from Open Philanthropy to Coefficient Giving, and P05, P06, P10, and P11 describe overlapping ecosystems.
- P07's donor suggestions are individual staff views, not institutional views.
- P03's catch-up growth analysis is a third-party or adjacent portfolio argument, not a CG institutional position.
- Open question: how much of the 2026 scale-up is AI and biosecurity versus other funds, and how much is Good Ventures versus other donors?

# Analyze the stated AI and biosecurity scaling rationale, urgency and short-timeline strategy.

## Question

**Deliverable:** a memo on why Coefficient Giving says it is going bigger and faster, together with the implications for grantmaker priorities and for the claims that could turn out to be wrong. The object of analysis is the stated rationale, not the merits of the underlying cause areas.

*Working standard: treat every rationale as an argument with load-bearing premises, a stated or implied falsifier, and a set of operational consequences. Where the sources disagree with each other, preserve the disagreement rather than averaging it. Where a claim is a projection dressed as a result, mark it as a projection.*

## Background

**Source base for this step:** P01 (Oehlsen, portfolio lead), P06 (CEO letter, explicitly framed as a personal reflection), P08 (Berger and Oehlsen, published under the former name Open Philanthropy), P03 (Oehlsen-adjacent growth/welfare modeling), D04 (AI Governance RFP), D02 (AI enablement), plus adjacent-funder comparators P10, P11, P12, D01, D03, and the S-series staff posts.

**Voice discipline.** *P06 is the CEO's personal reflection, not an institutional commitment; P01 is the GCR portfolio lead writing in an institutional register; P08 reflects a pre-rebrand institutional position. Merging these into a single institutional voice would erase the very differences that make the scaling rationale interpretable. Note also that the role brief and the grantmaker hiring posts are historical and do not establish a current vacancy.*

**Figure discipline.** "Committed" means grants formally recommended for funding; it is not the same as "conditionally approved," and it is not the same as published grants, because publication lags (P01 footnote 2). "On track to commit" for the current year is a projection. Critically, the two headline "over $1 billion" figures have different scopes and different years: P01's is AI safety, security and fieldbuilding projected for the current year; P06's is total organizational giving directed in the prior year. Conflating them inflates the AI story.

**Vocabulary fixed here.** Transformative AI (AI capable of driving changes at least as significant as the Industrial Revolution, per P08); TAIS = Technical AI Safety; absorptive capacity = the organizational ability to convert money into work; inframarginal = opportunities that clear the bar by a wide margin once created, as distinct from the marginal dollar that determines portfolio allocation (P06); and the detachment window = the period of accelerating frontier growth during which the frontier non-frontier GDP gap widens (P03).

## Sub-Question 1: 1. Decompose the rationale: two drivers and one enabling constraint

Reconstruct the scaling rationale as an argument. Separate the demand-side driver (capability progress and shortened timelines) from the supply-side driver (an expected wave of philanthropic funding), then identify the enabling constraint that actually determines the pace. State which premises carry the argument's weight.

### Step 1

#### Step 1.1: the demand-side driver.

P01 gives four sub-claims, and their evidential quality is not uniform. First, "many experts have updated toward thinking transformative AI may arrive sooner than they previously expected, potentially within the next few years," citing Cotra, Greenblatt and Clark. Second, the author's own timelines shortened. Third, the specific drivers of that update are given as rapid capabilities progress, the rise of agentic coding, unprecedented revenue growth at frontier AI companies, and the author's day-to-day use of AI. Fourth, most of the portfolio is claimed to be worth doing even under long timelines.

**The non-trivial move:** *the third sub-claim is not a forecast aggregate at all. "My own day-to-day use of AI" is first-person experiential evidence, and it cannot be aggregated with expert survey data because it is a sample of one, drawn from a person whose job depends on treating AI progress as consequential. A grantmaker rehearsing this rationale in an interview should present it as it is: a portfolio lead reporting an update, where the update is partly intellectual and partly experiential. The two are not the same kind of claim. Strong=true for the distinction, not for the claim.*

P06 complements this from the CEO's chair: short-timeline predictors have been more accurate than the author's own predictions over the last year, and the CEO has moved toward seeing continued fast progress as the default rather than an aberration. Two caveats from the same source must travel with that thought: the CEO calls out extrapolative bubble dynamics, and reports that METR's time-horizon measurement is struggling because models saturate the tasks. Saturation is ambiguous between a measurement artifact and a genuine ceiling, and the scaling rationale depends on which it is.

*P03 and D04 supply the catastrophic-harm framing, locating it "within the next 15 years" or "e.g. in the next 15 years." Note the rhetorical telescoping: the timeline horizon used to justify urgency varies from "the next few years" in P01 to "15 years" in D04. These are different decision environments. Under the first, only immediately deployable work clears; under the second, multi-year institution-building is rational. The scaling rationale quietly needs both, which is the first crack to note.*

**Step 2**

**Step 1.2: the supply-side driver, and why it is a real-options argument rather than a liquidity argument.**

P01 gives the supply-side argument in a sentence most readers skim: "Organizations need calendar time to expand, hire, and make plans, so resources received today are worth more than resources arriving in two years." The subtle logic is that this is not a claim about the time value of money. It is a claim about the growth rate of organizations under a time constraint. Capital cannot be converted into absorptive capacity faster than organizations can grow, so the marginal dollar today buys capacity that a dollar in two years cannot buy in time.

$$V(\text{dollar at } t = 0) > \frac{V(\text{dollar at } t = 2)}{(1+r)^2}$$

when the binding constraint is elapsed calendar time, not capital

*This formalization makes the falsifier visible. If organizations can absorb capital as fast as it arrives, the inequality collapses and waiting dominates accelerating. So the supply-side driver is logically downstream of the absorptive-capacity claim. That means a grantmaker evaluating the scaling rationale should probe the absorptive claim first, not the funding-wave claim.*

The same logic is restated at ecosystem level by an adjacent funder, P10: "Money gets absorbed by orgs that can hire people and run projects, and the field doesn't have enough of them." P10 also names the mechanism that makes front-loading valuable: the next 12 to 18 months disproportionately determine whether the ecosystem reaches meaningful size in time. Independent convergence on the same argument strengthens it, but note that P10 is a funder whose own proposition is to fund small early-stage projects, so it is not a disinterested witness to the claim that organizations are scarce.

P10 also supplies a live example of the same logic applied outside AI: the reported $825 million funding bump for GiveWell's recommendations for the coming year, up from a much smaller commitment, with the explicit reasoning that this is a one-off surge to fill cost-effective gaps now and create scalable opportunities for future donors. This matters because it shows the supply-side argument is a general posture at CG, not an AI-specific one — which is exactly what the falsification test needs, because a general posture is harder to reverse than a one-off.

**Step 3**

**Step 1.3: the enabling constraint, and the quantitative landscape it governs.**

P01 states the constraint plainly: "We remain capacity-constrained and the main thing preventing us from scaling our impact is staff time." This single sentence governs the interpretation of everything else. "Bigger and faster" is not primarily a statement about risk appetite or budget headroom. It is a statement that the binding resource is grantmaker judgment-hours, and that every operational change must be read as an attack on that constraint.

The reported giving series is the surface this constraint sits under. Treat the current-year figures as projected or on-track commitments.

Table 11: *Reported CG giving figures with their scope and status made explicit. The first and fourth rows are both "over $1 billion" and are not interchangeable.*

| **Series (source)** | **Prior year** | **Last full year** | **Current year (projected)** | **Status** |
|---|---|---|---|---|
| AI safety, security, fieldbuilding (P01) | $168M | $351M | over $1B on track | committed / projected |
| Biosecurity (P01) | $30M | $65M | more than $150M so far | committed / partial year |
| Technical AI Safety only (P06) | $46M | $157M | across 200+ grants in the last full year | committed |
| Total organizational giving (P06) | not given here | over $1B | targeting further scale | directed; different scope and year |

Multiples make the pace concrete. This is a trivial computation whose value is entirely in forcing the projected figures to be treated as floors rather than point estimates.

```
# Source figures in $M, as reported by P01 and P06.
# The current-year AI figure is reported as 'over $1B', so we treat
    it as a FLOOR.
ai_by_year = input_state["ai_giving_by_year"]      # {"2024": 168,
    "2025": 351}
bio_by_year = input_state["bio_giving_by_year"]    # {"2024": 30,
    "2025": 65}
tais_by_year = input_state["tais_giving_by_year"]  # {"2024": 46,
    "2025": 157}
ai_current_floor = input_state["ai_current_floor"] # 1000

def step_ratio(series):
```

```
    years = sorted(series)
    return series[years[-1]] / series[years[0]]

growth = {
    "ai_last_full_year_over_prior": step_ratio(ai_by_year),
    "bio_last_full_year_over_prior": step_ratio(bio_by_year),
    "tais_last_full_year_over_prior": step_ratio(tais_by_year),
    "ai_projected_over_last_full_year_floor": ai_current_floor /
    ai_by_year["2025"],
}
output_state["growth_multiples"] = growth
```

*Read the outputs qualitatively: the AI portfolio roughly doubled then is projected to more than double again; TAIS grew fastest from the smallest base; biosecurity grew from a much smaller base but at a comparable multiple. The important interpretive point is that a 3x-plus budget increase against a fixed or slowly growing staff count is arithmetically a demand for roughly 3x the dollars moved per staff-hour. That is the operating meaning of "going bigger and faster," and it is what makes step 2 the real content of the strategy.*

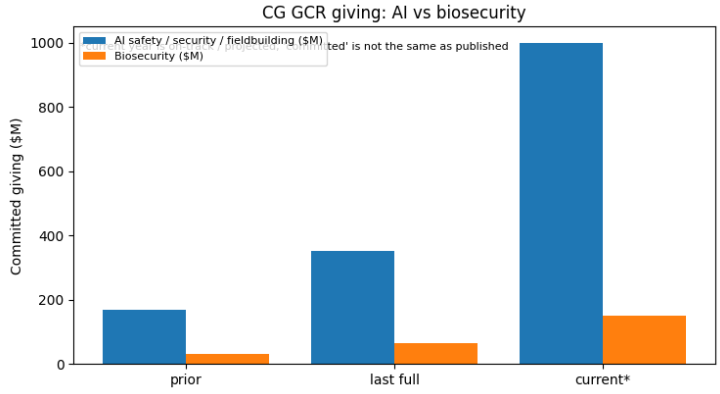


Figure 1: Graphical analysis for Step 3.

### Step 4

**Step 1.4: name the load-bearing premises, and name the tension that the rationale has to survive.**

- **Timeline premise:** transformative AI is near enough that immediate work dominates deferred work. Falsifier: continued fast progress turns out to be an extrapolative bubble, or capability measurement saturation reflects a genuine ceiling rather than a task-suite artifact.

- **Absorptive premise:** organizations cannot absorb capital as fast as it arrives, so capacity built now is worth more than capacity funded later. Falsifier: the small-grant niche turns out to be adequate and a funding wave finds ready homes.
- **Supply premise:** a wave of philanthropic capital is genuinely coming, not merely forecast. Falsifier: AI wealth converts to consumption and equity rather than intended philanthropic spend, or conversion is much slower than assumed.
- **Capacity premise:** CG can build enough grantmaker capacity to deploy at the projected scale. Falsifier: the organization cannot recruit against lab compensation, and dollars per staff-hour rise faster than the process can absorb.

**The tension:** *P01 says CG expects to keep saying no to many proposals, even good ones, because the goal is impact and not moving money. Yet the supply-side driver says the ecosystem must be seeded now because organizations are scarce and a wave is coming. Saying no to good proposals while funding the creation of new organizations is coherent only if CG believes it can create better opportunities than it can find. That belief is a real premise, and it is the one the next subquestion examines day to day. Disagreement to preserve: P10 argues the opposite emphasis at the small end, treating breadth and tolerance of failure as the way to build absorptive capacity.*

## Sub-Question 2: 2. Bigger, faster, proactive: one capacity strategy in three faces

The listed changes — larger grants, faster approvals, proactive seeding — read as three separate initiatives. Argue that they are better understood as three faces of one strategy aimed at the binding constraint identified earlier, and identify the costs and risks each face introduces.

### Step 1

#### Step 2.1: larger grants are an arithmetic attack on the staff constraint.

P01 reports that the majority of GCR grants over $20 million ever made were made in the current year, and that where conviction is high CG wants to bet big from the start, citing $160 million for Resolution. The obvious reading is higher risk tolerance. The less obvious reading is that grant size is a lever on dollars moved per unit of grantmaker time. If your constraint is staff-hours, the cheapest way to raise throughput is to raise the average ticket size, because the fixed cost of an investigation is roughly independent of the grant amount.

```
# Illustrative arithmetic. The staff-hour figures are INVENTED
    placeholders chosen
# only to display the shape of the tradeoff, not estimates from any
    source.
```

```
grant_sizes = input_state["grant_sizes"]        # {"small": 100_000,
    "mid": 2_000_000, "large": 160_000_000}
invest_hours = input_state["invest_hours"]      # illustrative
    investigative staff-hours

dollars_per_hour = {k: grant_sizes[k] / invest_hours[k] for k in
    grant_sizes}
output_state["dollars_moved_per_staff_hour"] = dollars_per_hour
```

*The result is that dollars per staff-hour spans three orders of magnitude across grant sizes. Even allowing large error in the illustrative hours, the ordering is robust: raising average ticket size is a far more powerful throughput lever than shortening investigations. This reframes a claim that otherwise sounds like pure appetite. The memo should say plainly that bigger grants are simultaneously an impact bet and a capacity strategy, and that these two justifications can come apart: if a large grant is made mainly for throughput, the quality bar and the capacity motive are in tension, and only one of them is being served.*

P01 also reports CG encouraging grantees to raise their ambitions, including larger compute budgets and higher salaries. This is the mirror image of CG's own constraint, and it introduces a principal-agent caution that the source itself acknowledges: "we don't think every organization or idea should scale." Encouraging absorption is only correct when the grantee's marginal dollar is genuinely high-return; otherwise the encouragement is a budget-inflation vector.

### Step 2

#### Step 2.2: faster approvals are a delegation-of-judgment move, and they raise the value of transparent writing.

P01 reports that a typical grant now needs active sign-off from only one person in addition to the primary grant investigator, that steps adding delay were cut, especially for grants clearly above the bar, and that the $10 million institute grant to Tsimerman was approved twelve days after the first conversation.

*The non-trivial interpretation is about information location, not about standards. Shortening the ladder to one extra sign-off is a claim that the primary investigator holds the most context and that additional reviewers add less than they cost. If that claim is right, speed and quality improve together. If it is wrong, error concentrates in a single reviewer's blind spots, and the system loses exactly the error-catching that multiple reviewers provide. The design therefore shifts quality assurance upstream, into how the investor gathers and writes evidence.*

This is why reasoning transparency stops being a stylistic virtue and becomes an operational requirement. D01 prescribes opening with a linked summary of key takeaways, indicating which considerations are most important, and stating confidence and kinds of support for major claims. In a one-reviewer work-

flow, those are precisely the affordances that let a second person decide quickly whether the decisive considerations were handled. A grantmaker should treat D01 compliance as load-bearing infrastructure for the scaling strategy, not as good manners.

**Caveat about evidence:** *a single 12-day approval is anecdotal. It is a proof of concept, and public write-ups are subject to survivorship: fast approvals that worked are announced; fast approvals that misfired are not the sort of thing an organization publishes. A grantmaker reasoning from this evidence should say so rather than treating the example as validation of the process at portfolio level.*

**Step 3**

**Step 2.3: proactive seeding is the highest-variance face, and the one that interacts hardest with the no-to-good-proposals policy.**

P01 describes proactive seeding as part of the job: when an organization should exist and does not, CG considers helping create it, with Project Tailwind aimed at finding and funding would-be founders, and a footnote preserving the tradeoff that broad seeding yields to prioritization of the highest-impact work when the two conflict.

*The deeper logic is in P06, which distinguishes marginalist funding from inframarginal bets. Marginalist reasoning allocates dollars to equalize expected returns within and across funds. An inframarginal grant is one that clears the bar by a wide margin once it exists, but only because someone recruited a founder or effectively incubated the organization first. Such grants are invisible to marginalist selection because there was nothing to select until someone did the labor. P06 states the tradeoff precisely: the ownership required to enable these grants competes for the same scarce grantmaker time as optimizing the marginal dollar.*

A concrete instance of the pipeline this produces, from an adjacent funder rather than from CG: P11 reports that an initial $143k regrant seeded Timaeus, which later merged into Resolution and received a $160 million grant. The non-trivial point is not the multiplier, which is retrospective and winner-selected. It is that the large grant at the top of the funnel is only available because a small grant at the bottom existed years earlier. This is the strongest available support for CG's supply-side argument, and it also shows how the ecosystem's segmentation works: CG operates the large end, while P10, P11 and D03 operate the small end.

**Preserved disagreement:** *P06 itself warns against the strategic-funder trap of treating grantees as subcontractors for a funder's vision, and notes that the virtue of marginalist funding is that it is relatively strategy-agnostic, which facilitates openness to grantees' local knowledge. Proactive seeding pulls in the opposite direction. P06 explicitly refuses to resolve this, saying the right balance varies across focus areas and individuals. A memo that resolved it would be misrepresenting the source.*

## Sub-Question 3: 3. Short timelines, hedged portfolios, and a posture that looks like a reversal

Reconcile three statements that appear to conflict: CG is planning for short timelines specifically, most of its work is worthwhile even under long timelines, and AI giving may not increase dramatically in the current year. Explain how the three can all be true, and what the result implies for a grantmaker's planning assumptions.

### Step 1

**Step 3.1: the short-timelines team is a portfolio-of-teams move, not a change in any single team's median assumption.**

P01 says a team was launched to work out what to do if transformative AI arrives within the next few years, and that this work complements other AI teams which are organized thematically and hedge across different possible timelines. P06 names the lead and describes the work as building a team and triaging projects.

**The non-trivial reading:** *the organization gets timeline diversification from the ensemble of teams, not from any one team's forecast. That means an observer cannot infer CG's central timeline assumption from the existence of the short-timelines team, and cannot infer that the thematic portfolios have shifted their median assumptions either. The statement "CG is taking short timelines seriously" is compatible with the main portfolios continuing under longer-horizon assumptions. Conflating the two is the most common interpretive error available here.*

### Step 2

**Step 3.2: robustness across timelines and the dedicated tail hedge are consistent because they answer different questions.**

*"Most of our work is worth doing even under long timelines" is a robustness claim about a broad portfolio: the portfolio does not crumble if timelines extend. "We are planning for short timelines specifically" is a claim that the portfolio is incomplete without a component that only pays off in the short-timeline branch. Robustness across worlds and insurance against one world are complementary, not contradictory. In decision terms, the general portfolio maximizes expected value under a wide prior, and the dedicated team adds a component whose value is concentrated in the tail.*

P08 supplies the underlying situational assessment on which any timeline branch rests, and it is worth isolating because it is a quantified institutional claim with a date: an estimate made in 2016 that there was at least a 10% chance of transformative AI arriving within 20 years. The source then observes that what felt bold at the time now looks accurate or conservative in retrospect, and that many staff and external experts put the odds by 2035 considerably higher. That

trajectory of estimate revisions is itself a fact that bears on the scaling rationale, and it is also the kind of fact that can be reinterpreted after the fact by a lucky regime, which is exactly the caveat P06 raises about extrapolative expectations.

**Step 3**

**Step 3.3: the current-year posture diverges from the headline narrative, and a grantmaker should plan for the posture.**

*P06 says the organization's AI giving may not actually increase dramatically in the current year, and that if increasing funding for existing grantees competes with creating new highly absorptive organizations, it would prefer to prioritize longer-term growth. This sits in the same document as the ambition to scale, and it is the single most decision-relevant statement for a new grantmaker.*

Why the two are not a contradiction: the headline describes a multi-year trajectory, while the near-term posture reflects a specific tradeoff in which budget increases were already committed across funds and the marginal value of another dollar of near-term throughput is lower than the marginal value of building future absorptive capacity. P06 states the general form of this shift explicitly: with nearly every fund working with a higher budget, attention should move away from marginal tradeoffs and toward more ambitious goals, and many of the best grants have been deeply inframarginal.

**Implication for a grantmaker, stated as a planning assumption:** *do not assume budget headroom creates spend pressure. Assume instead that the scarce resource is the ability to identify or create inframarginal opportunities, that a flat-to-modest near-term AI budget is a live possibility, and that proposals which build future capacity without moving large sums now may be favored over proposals that move large sums now without building capacity. A candidate who rehearses only the growing-numbers story will misread the actual job.*

## Sub-Question 4: 4. Biosecurity: a parallel rationale with a different evidentiary base

Show why the AI scaling argument does not transfer intact to biosecurity, despite both portfolios scaling rapidly. Identify the drivers that are actually doing the work in biosecurity, and what that implies for a generalist grantmaker moving between the two.

**Step 1**

**Step 4.1: only one of the AI urgency drivers is available to biosecurity.**

The AI urgency drivers cited in P01 are capabilities progress, agentic coding, frontier revenue growth and personal use. None of these is a biosecurity phenomenon. The biosecurity portfolio nevertheless grew from $30 million to $65 million to more than $150 million across the reported years, and the supply-side

argument applies to it in exactly the same form, since the funding wave and the calendar-time argument are cause-agnostic.

What supplies the demand-side urgency in biosecurity is the intersection with AI. P08 lists extreme misuse as a concern pathway in which AI could help bad actors engineer far more lethal and transmissible pathogens, and notes that current systems have rapidly improved at virology, which could yield scientific breakthroughs but could also pose future biological risks. P08 also reports, as evidence that the concern is not merely theoretical, that two funded studies pointed in opposite directions on whether current language models materially increase biological attack risk, with both produced by groups the funder supported.

**The non-trivial consequence:** *biosecurity urgency is largely a function of AI urgency through one specific channel, so if the AI capability premise weakens, the biosecurity portfolio loses part of its demand-side case while keeping its supply-side case. A grantmaker should be able to state this asymmetry rather than presenting both portfolios under a single timeline story.*

**Step 2**

**Step 4.2: the second driver is individual ownership, which is a personnel argument rather than a portfolio argument.**

P06 names the biosecurity lead as an exemplar of taking full responsibility for a problem, and cites work on reducing risks from mirror bacteria and a four-pillared plan to avoid engineered biological catastrophes: personal protective equipment such as elastomeric respirators, biohardening of buildings and infrastructure, pathogen-agnostic detection systems, and rapid medical countermeasures. P04 gives the same person in first-person form, describing a grant created because the grantmaker went looking for it, funded with a couple of reference checks after one conversation.

*The interpretive point: the biosecurity scaling rationale is substantially a story about an unusually high-ownership individual general manager who found and created opportunities, and only secondarily a story about a forecast-driven imperative. That makes it replicable in principle but fragile to a single departure, and it makes the "general manager versus marginalist" tension from step 2 concrete rather than abstract. P06 is explicit that it wants the ambition and ownership of general managers without losing the curiosity and humility of marginal funders, and that the balance will vary across focus areas.*

P04 also supplies the operational texture a generalist needs: a typical biosecurity grantmaker may be investigating six to twelve grants at various levels of focus, a meaningful share of time goes to field building such as fellowships and programs, another large chunk to finding individuals who could contribute a great deal but have not yet made the jump, and the team runs an RFP. This is a different workflow from making a small number of enormous technical bets.

**Step 3**

**Step 4.3: do not let the AI portfolio's intuitions supply the biosecurity portfolio's judgment.**

- **Bottleneck type:** the technical AI safety team reported that its bottleneck shifted from research paradigms to funding and talent after capabilities made empirical safety work more tractable. The biosecurity description in P04 points to talent and field-building bottlenecks, and to individual transitions. The bottleneck diagnosis determines whether money or people are the constraint.
- **Grant archetype:** AI includes nine-figure single-organization bets. Biosecurity in P04 includes a grant of several months of full support so one person can leave a job and transition into the field. Both are proportionate due diligence at their respective scales, and importing a large-bet mindset into a small-transition decision would be a category error.
- **Evidence base:** *AI urgency rests on capability measurements and forecast aggregates. Biosecurity urgency rests on a smaller and more contested literature, as P08's account of two opposing funded studies illustrates. The grantmaker should expect the evidentiary floor to be lower and the political economy more constrained.*

Adjacent evidence on why the political economy is constrained comes from a different corner of the corpus: N14 documents a series of policy proposals that foundered because there was no expert consensus on the problem, and argues that initiatives which build on a majority expert view are not useful unless such a majority already exists. Biosecurity risk governance is exactly the type of area where consensus is incomplete, which is a structural headwind that AI-specific scaling enthusiasm does not encounter in the same form.

## Sub-Question 5: 5. Implications for grantmaker priorities, with calibrated answers

Translate the rationale into operating priorities for an evaluator or grantmaker, then produce calibrated answers to the interview question "why is Coefficient Giving going bigger and faster?" and at least two defensible answers to the crux question of whether the scaling is appropriately sized.

**Step 1**

**Step 5.1: the operating priorities that follow from the analysis.**

- **Treat staff time as the binding resource.** The highest-value grantmaker activity is often creating an inframarginal opportunity rather than evaluating a marginal inbound one. Expect to spend scarce time recruiting founders, incubating, and shaping proposals rather than only reading them.

- **Write for a single revisor.** Under the shortened approval ladder, the memo is the quality gate. Front-load takeaways, mark which considerations are decisive, state confidence and the kind of support behind each major claim, and say what would change your mind.
- **Scale diligence to stake, and be able to defend the bar.** D04 states that smaller proposals should ideally be funded elsewhere and that first-time grants typically begin in a specific band. When your own portfolio includes both small transitions and enormous bets, the proportionate-diligence argument must be made explicitly rather than assumed.
- **Do not read budget growth as spend pressure.** *The near-term posture may favor capacity-building over throughput. Saying no to good proposals is part of the job, and the reason should be stated in terms of the highest-impact alternative, not in terms of discomfort with the proposal.*
- **Use the tools the organization actually built.** D02 describes a grant copilot that gathers information and drafts write-ups, a back-of-the-envelope calculator generator, and multi-agent research workflows, motivated explicitly by staff time as the key bottleneck. The AI-enablement posture and the scaling rationale are the same argument seen from the operations side, and D02 also cautions that generated prose is poor and that automation can create an illusion of speed.
- **Expect the ecosystem split to shape your mandate.** D03 and P10 describe small early-stage funders operating below the threshold at which CG prefers to work. A grantmaker should know when the right recommendation is to route a proposal elsewhere rather than to fund it.

**Step 2**

**Step 5.2: a calibrated answer script, with its assumptions and limits exposed.**

**Script, roughly 45 seconds:** "As I read the published material, there are two independent drivers and one enabling constraint. The demand side is shortened timelines and rapid capability progress, which the sources attribute to capabilities, agentic coding and frontier revenues, and which they also flag as uncertain. The supply side is an expected wave of philanthropic funding, combined with the argument that organizations need calendar time to grow, so a dollar today buys capacity a dollar in two years cannot buy in time. The enabling constraint is staff time: the stated main limit on impact is grantmaker capacity, not dollars. That is why bigger grants, faster approvals and proactive seeding are best understood as one strategy aimed at moving more dollars per staff-hour rather than as a general rise in risk appetite. Two caveats I would want to keep visible are that current-year figures are commitments on track rather than realized disbursements, and that the sources themselves say AI giving may not rise dramatically this year and may prioritize longer-term capacity over near-term

throughput."

**Limits of the script:** *it does not claim to know the organization's internal weighted view on timelines, it does not treat any single year's figure as a trend, and it does not merge the CEO's personal reflection with institutional commitments.*

**Step 3**

**Step 5.3: three defensible answers to the crux question of whether the scaling is appropriately sized.**

1. **Answer A, appropriately sized.** The binding constraint is capacity rather than capital, so front-loading is justified by the calendar-time argument and is robust across a wide range of timeline assumptions; the ecosystem's segmentation, with large high-conviction bets at this end and small early-stage bets elsewhere, is a sensible division of labor. What would move me: evidence that organizations can absorb capital as fast as it arrives, or that the funding wave is not materializing.
2. **Answer B, under-scaling at the small end.** The small-grant niche is thin, tolerant of failure, and hits-based, and the same sources argue that organizations are the scarce input while individuals' first grants are some of the highest-return. A posture that keeps saying no to good proposals, even for good reasons, may fail to build the absorptive capacity the supply-side argument says is needed. What would move me: evidence that small grants are being crowded out, or that the seeded organizations are not reaching follow-on funding.
3. **Answer C, premature or over-levered.** If continued fast progress is an extrapolative expectation, or if capability measurement saturation reflects a genuine ceiling, then pre-seeding organizations ahead of a funding wave creates entities that the wave never funds. The sources themselves supply this risk: the bubble-dynamics caveat and the warning against the strategic-funder trap. What would move me: a funding wave that arrives but routes to existing organizations, leaving newly seeded ones unsupported.

**Marginal note on answer quality:** each answer states what evidence would change it. That is the property that distinguishes a rehearsed position from a defended one, and it is also what the sources ask of their own grantees when assessing strategic judgment.

## Sub-Question 6: 6. Falsifiers, discounts on headline numbers, and how a grantmaker should update

Close the memo by separating what the sources support from what they project, listing the falsifiers that would change a grantmaker's priorities, and specifying

the exercise that would test whether the interpreter understands the argument rather than merely reproducing it.

**Step 1**

**Step 6.1: separate the three tiers of claim, and apply the discount.**

Table 12: Three tiers of claim with an explicit discount. The fourth row is the inference a reader is most likely to smuggle in without noticing.

| **Claim** | **Tier** | **Discount to apply** |
|---|---|---|
| Prior-year giving figures ($168M, $351M, $46M, $157M, $30M, $65M) | source-supported commitment data | Minimal, but remember these are commitments, not disbursements, and publication lags. |
| Current-year figures ("over $1B", "more than $150M so far", August's $352M) | on-track projection or partial-year actual | Treat as floors and as incomplete; do not annualize a partial year. |
| Shortened timelines and the funding-wave estimate | attributed expert view plus the author's own update; adjacent-funder forecast | State as a forecast with the source's own hedges attached, including the bubble caveat. |
| Causal claim that scaling causes impact | inference, not a source-supported finding | *Do not assert it; the sources assert scale and speed, not that scale and speed produced the impact.* |

*One further discount specific to voice: several of the sharpest statements about the pace of AI progress come from a personal reflection and from a portfolio lead's first-person update, and one of the quantified estimates in P08 predates the rebrand and is described in that source as having felt bold at the time. These are still evidence about the organization's reasoning, but they are weaker evidence about its commitments.*

**Step 2**

**Step 6.2: the falsifier list, along with the priority shift each would trigger.**

- **If capability progress flattens or measurement saturation proves real:** shift weight from short-horizon, immediately deployable work

toward durable institution-building, and raise the required evidence for urgency-based prioritization.

- **If the funding wave stalls:** the calendar-time argument loses force, pre-seeded organizations become unfundable, and the rational posture moves toward funding existing, proven absorbers rather than creating new ones.
- **If the wave arrives but routes to incumbents:** the seeding strategy has failed on its own terms even though the headline rationale is vindicated; the grantmaker should prefer proposals with realistic follow-on funding paths.
- **If staffing cannot scale against lab compensation:** the binding constraint binds harder, and the organization is pushed further toward high-ticket grants and automation, with a correspondingly higher concentration of judgment in fewer reviewers.
- **If a fast, large approval misfires publicly:** expect a partial reversion to more review for large grants, and expect the reasoning-transparency standard to be enforced more strictly rather than relaxed.

**Step 3**

**Step 6.3: the exercise that tests comprehension rather than recall.**

Exercise: identify which scaling claims are load-bearing and what evidence would change a grantmaker's priorities. Concretely, write one page that (a) names the four premises from step 1.4, (b) for each, states one observation that would materially weaken it, (c) notes which premise's failure would most change your own allocation of attention as a grantmaker, and (d) states what you would do differently the following week if that premise weakened.

*The test of a good answer is the direction of dependence. A weak answer names a premise and vaguely says it would matter. A strong answer says, for example: if the absorptive premise weakens, proactive seeding becomes low-value relative to funding existing absorbers, so I would rebalance my week toward renewals and away from incubating; if the supply premise weakens but the absorptive premise holds, the right move is the opposite, because scarcity of organizations plus scarcity of capital means the marginal dollar is unusually valuable and throughput should rise.*

*One closing observation that the sources support but do not state: the scaling rationale is internally consistent only if the organization can keep doing two things at once — moving more money per unit of staff time, and finding or creating opportunities that justify the larger tickets. The first is an operations problem with a partly solved solution set in D02. The second is a judgment problem with no solution set, only a culture. That asymmetry is the most useful*

*thing to carry into an interview, because it locates the role precisely: the role exists to supply the second thing.*

# Reconstruct the day-to-day grantmaker workflow from investigation to approval.

## Question

Reconstruct the day-to-day grantmaker workflow at Coefficient Giving from the point at which an opportunity enters the pipeline to the point at which a grant is approved and managed, and produce a process map showing the investigation, the evidence-seeking, the write-up, the team-lead review, the approval step and the post-approval relationship.

The map must name the **decision points** at which the workflow could go differently, and the **bottlenecks** that actually constrain throughput. It must distinguish what the sources state from what is being inferred, and it must survive the observation that the public evidence on this workflow is thin: a three-person recruiting post, a scaling announcement, and a published account of internal AI tooling.

## Background

The attributed role brief states the competencies the workflow is meant to exercise: comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, judging people and organizations, and communicating concisely, calibrated and transparently. The workflow is therefore the concrete site at which those abstractions become observable behaviour, and an interview that probes the workflow is probing those competencies indirectly.

The primary evidence is **SOURCE P04**, a day-in-the-life post covering three grantmakers across AI governance and policy, infosecurity and biosecurity, published as a recruiting post with a historical deadline (May 17). It is a three-person sample, chosen for representativeness by the communications team rather than sampled. Supporting evidence comes from P01 (shortened investigations, approval pushed to grantmakers, the Tsimerman grant approved twelve days after the first conversation), P06 (a redesigned approval process for a technical-AI-safety RFP, and the internal AI-adoption programme), D04 (the AI governance RFP's process design and its capacity problem), D02 (internal AI tooling built for grantmaking), P05 and P02 (what an open call buys) and P12 (trust as the deployment bottleneck).

Three caveats govern the whole reconstruction. First, no source publishes an internal process document or an investigation template; every diagram below is a reconstruction from narrated examples, and should be labelled as such. Second, the sources describe a process that is deliberately changing, so a snapshot is a snapshot of a moving target. Third, the source material is self-interested in the specific sense that it is written for external readers and prospective applicants, which biases it toward describing the process as faster and more legible than it

feels from inside.

## Sub-Question 1: Entry Points: How an Opportunity Enters the Pipeline

The attributed brief compresses the front of the workflow into a single verb: grantmakers **investigate funding opportunities**. The corpus shows at least two structurally different processes behind that verb, and the difference is not cosmetic: it changes where the bottleneck sits, who performs the work, and how counterfactuality is assessed.

Reconstruct both entry flows, examine what an open call actually buys the funder beyond a set of fundable proposals, and then examine proactive origination and the asymmetry it creates.

### Step 1

Inbound: the opportunity arrives as an artifact someone else wrote. Trevor Levin's representative Tuesday contains an hour-long call with an applicant for a large grant, preceded by teammates' review questions (SOURCE P04). The AI governance RFP runs an Expression of Interest stage, deliberately short (about one hour of applicant time, per SOURCE D04), before a full proposal of two to six hours and then a grant investigation. The Effective Careers RFP received 474 applications and moved 111 to a second stage (SOURCE P02).

Outbound: the opportunity does not exist yet and the grantmaker creates it. Chris Bakerlee learned of a master's graduate in Europe through a contact, had one conversation and two reference checks, and recommended several months of full support so the person could leave their job; he notes that the grant **would not have happened if I hadn't gone looking for it** (SOURCE P04). Abbey Chaver describes scouting for talent and projects and trying to sell them on starting a project (SOURCE P04). Emily Oehlsen reports brainstorming more than 200 ideas her team would be excited to fund if a promising founder existed, and offering pre-seed, seed and scale funding tiers (SOURCE S09).

Table 13: Both flows converge on the same investigation and sign-off chain; only the front of the process differs. The convergence is an inference: the sources describe them separately and do not draw the merged pipeline.

| **Dimension** | **Inbound: open call or application** | **Outbound: scouted, seeded or recruited** |
|---|---|---|
| Trigger | A public RFP or an unsolicited proposal (SOURCE D04) | A mapped gap, a person the grantmaker wants to recruit, or a founder search (SOURCES P04, S09) |

| Dimension | Inbound: open call or application | Outbound: scouted, seeded or recruited |
|---|---|---|
| First artifact | An Expression of Interest, one suggested hour of applicant time (SOURCE D04) | A conversation or an internal memo; there may be no written proposal until the grantmaker elicits one (SOURCE P04) |
| Who does the front-end work | The applicant writes; the funder triages | The grantmaker writes, recruits and co-designs |
| Characteristic bottleneck | Review capacity at volume: the number of applications has overwhelmed staff capacity (SOURCE D04) | Grantmaker hours and trust with a founder who does not yet exist (SOURCES P06, P12) |
| Counterfactuality test | Must be estimated: would another funder or another donor have backed this? | High by construction; the failure mode flips from deadweight to hubris, the risk that the funder's own model of what should exist is wrong (SOURCE S08) |

The non-trivial consequence is that in outbound mode the grantmaker **co-authors the object of evaluation**. Chaver states the reframe explicitly: she does not **just consider whether a grant clears the bar, but whether I can amplify its impact by identifying the right goal or bringing in key talent** (SOURCE P04). If that is the job, then the investigator is not a scorer applied to a fixed proposal; the proposal is a variable the investigator controls. That is a strong claim about the role, and it is in direct tension with the marginalist framing that the CEO says he wants to preserve, where if an opportunity clears the cost-effectiveness bar, you should fund it (SOURCE P06 / S08).

**Step 2**

An open call is not only a sourcing mechanism. P05 lists four things a round buys: attracting co-funding, mapping the landscape more systematically, comparing opportunities side by side, and surfacing organizations **we wouldn't otherwise have seen**. The fourth is the epistemically interesting one, and P02 supplies evidence that it worked: the team **systematically underestimated both the size and the quality of the effective-careers space** and **mispredicted its shape**, having put a 25 percent chance on receiving more than 40 applications and receiving almost 500.

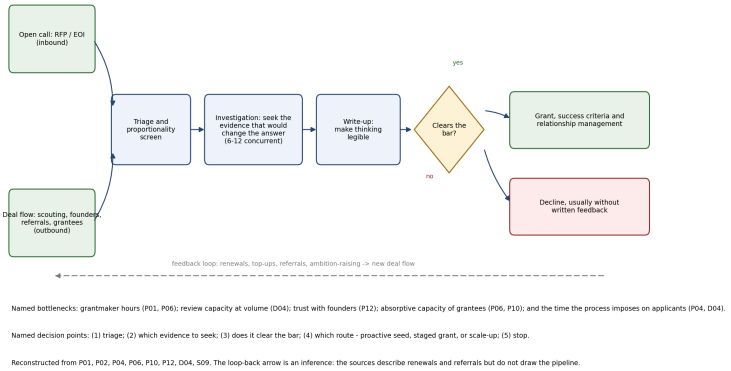


Figure 2: Graphical analysis for Step 1.

The non-trivial reading is that the RFP is a **measurement instrument on the funder's own model of a field**, and the forecast error is part of the return. But the instrument burns the resource it is meant to allocate. D04 documents the consequence at the AI governance RFP: rather than close the round, staff **raised the bar for which EOIs and full proposals we invite forward** so that proposals previously invited would now be rejected on capacity grounds. The design question that follows is a cost-per-unit-of-information question: what did this round buy, in mapped opportunity, per grantmaker-hour consumed.

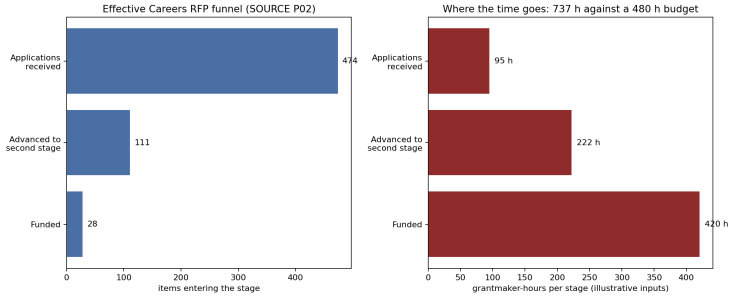


Figure 3: Graphical analysis for Step 2.

### Step 3

The largest grants in the public record were largely originated rather than received. P01 reports 160 million dollars for Resolution, the largest AI safety grant ever made, supporting a new organization led by Geoffrey Irving, and ten million dollars to help a Fields Medal recipient establish an institute, approved

twelve days after the first conversation. S09 reports a founder call with pre-seed, seed and scale tiers. S03 reports incubator participants raising pre-seeds above 500 thousand dollars from Coefficient Giving, Longview and BlueDot.

The non-trivial asymmetry: in outbound mode counterfactuality is high by construction, so the usual question, would this have happened anyway?, is answered before it is asked. P05 confirms that the Effective Giving portfolio applies **quality and counterfactuality adjustments** to reach a five-times return on investment, and S10 names the inbound failure modes directly: funding based on cost rather than value delivered, a game of chicken between funders producing glacial response times, unclear credit allocation and double-counted impact. Those are all inbound problems. The outbound problem is the mirror image, and Alexander Berger names it himself: he wants to **avoid the common strategic funder trap of thinking we have all the answers and just need to slot grantees into our vision, almost like subcontractors** (SOURCE P06, echoed in S08).

For an interview, this is where a candidate can differentiate. Saying I would investigate opportunities thoroughly is consistent with both flows and therefore says nothing. The stronger move is to say which flow the opportunity came through, which counterfactuality test applies, and which failure mode that flow is prone to. P05 lists the unresolved version of this as an open question in its own strategy: ***How much should we shape this ecosystem versus respond to it?***

## Sub-Question 2: The Investigation Itself: Proportionate Effort and the Evidence That Would Change Your Mind

The brief asks for **proportionate due diligence**. Proportionate to what? The corpus contains one operational answer stated by a working grantmaker, one project-management frame with an explicit scope condition, and one structural fact about how a grantmaker's attention is divided.

Reconstruct the evidence-seeking core, formalize the rule that governs it, and identify where the frame breaks down.

### Step 1

Trevor Levin states the rule: **ask, early and often, what piece of evidence would be most likely to change my mind, and to seek that evidence out** (SOURCE P04). He gives two instantiations. If a proposal looks weak but sits in an unfamiliar area, and the right expert vouching for it would change his mind, he schedules a fifteen-minute call with that expert. If the largest risk is a project lead who seems spread too thin, he has a conversation with them to get a read on how deeply they are thinking about failure modes. The goal, in his words, is **the most decision-relevant information as time-efficiently as possible, while avoiding unnecessary work for the organizations**

**you're evaluating**.

That last clause is the non-trivial part. It converts due diligence from a search problem into a **two-sided cost problem** in which the applicant's hours are a real cost. D04 makes this explicit at the criteria level: the cost-effectiveness assessment **also takes into account any additional costs the project may impose on others, such as the time required from evaluators or collaborators**. A stop rule that counts only the funder's hours is therefore wrong on the funder's own stated criteria.

$$P(E) = p(E \mid +)\,p(+) + p(E \mid -)\,p(-), \qquad P(+ \mid E) = \frac{p(E \mid +)\,p(+)}{P(E)}$$

Formalizing the rule this way yields a usable output: the break-even burden the funder may impose on a counterparty before a short call stops being worth it. Under the illustrative inputs below, the call is worth making until it costs the counterparty roughly five hours. The number is invented; the **form of the calculation** is the transferable part, and it is exactly the kind of reasoning the brief's reasoning quantitatively and probabilistically about expected value is pointing at.

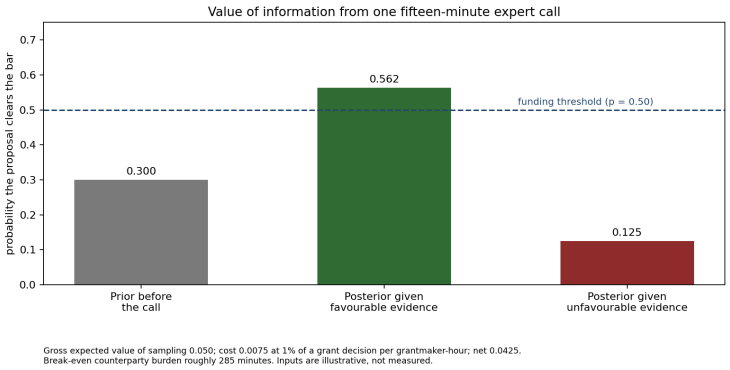


Figure 4: Graphical analysis for Step 1.

### Step 2

Abbey Chaver describes her operating frame, imported from product management: what work must be done to reach the goal, what dependencies exist, which stakeholders must be bought in, and how success will be measured (SOURCE P04). The frame is real practice, and it is a competent transplant of an execution discipline into grantmaking.

The scope condition is stated in the same paragraph and is easy to miss: **This is especially true in security, where many problems are straightforward**

**execution challenges**. The frame assumes that the binding constraint is execution. It is therefore weakest precisely where the field's hardest disagreements live, in what should exist at all. P08 supports this reading: from the beginning the funder **funded many grantees who disagreed about both the nature of risks posed by AI and the best ways to mitigate those risks**, and calls that a feature rather than a bug. Where the disagreement is about the objective, a dependency-and-milestone frame cannot adjudicate it, because it presupposes the objective. A candidate who applies it universally signals that they have not noticed the condition.

A second, subtler point sits in the same source. Chaver does not only score proposals; she tries to **amplify impact by identifying the right goal or bringing in key talent**. So the investigation has two outputs: a decision, and a modification to the thing being decided about. Most grant-evaluation heuristics assume the second output does not exist.

**Step 3**

Two facts about attention are stated plainly and are easy to skim past. Chris Bakerlee is **investigating 6-12 grants at various levels of focus** at any given moment. Chaver splits her time **more or less equally between meetings, research, and writing grant investigations or memos** (SOURCE P04).

The non-trivial consequence is that the workflow contains a **scheduling problem nested inside a judgment problem**, and the two interact. Which evidence a grantmaker seeks is determined by how much attention each investigation has been allocated, and that allocation is itself made on the basis of a first impression. Attention and judgment are not separable, which is why a candidate's claim that they would investigate everything thoroughly is not merely infeasible but incoherent.

Second, the modal outcome of the workflow is rejection. D04 states that the funder is **not planning to provide feedback for EOIs in most instances**, and P01 states that the funder expects to **keep saying no to many proposals, even good ones**. This sits in visible tension with the norm of reasoning transparency documented in D01, which recommends indicating confidence and support for major claims. The tension is not a defect to be scored against the institution: D01 is addressed to published analyses read by the public, while declines are private decisions read by nobody. But the candidate should be able to name the tension and say where transparency is owed and where it is not, which is a more discriminating answer than either they should explain everything or no explanation is owed.

## Sub-Question 3: Write-Up, Review and Approval: Where the Bottlenecks Actually Sit

The tiny deliverable asks for the review and approval chain with its bottlenecks and decision points. The corpus supplies an approval topology, a statement about what the write-up must accomplish, and enough about staffing economics to identify what is actually scarce.

Work through the artifact first, then the approval structure, then the bottleneck analysis, and finally the place of AI tooling in each.

### Step 1

Levin spends roughly two hours finishing a grant write-up, which he describes as **organizing the information and making my thinking legible to decision-makers** (SOURCE P04). The phrase is doing real work. The artifact is not a report about the world; it is an instrument for a reader who was not in the room and who must now sign something.

D01 specifies the top three requirements for such a document: open with a linked summary of key takeaways; indicate which considerations are most important to those takeaways; and indicate how confident you are in major claims and what support you have for them. It also lists the kinds of support that distinguish an expert call from a skim from a general intuition, and specifies that confidence can be expressed in words, probabilities or intervals depending on what the claim is worth.

L02, a local craft document in the corpus, supplies the sequencing discipline that D01 leaves implicit: meet the reader where they are, take them to the claim, show how you could be moved, show the evidence, show whether the evidence moved you, and show where they should now be. The two are complementary, and the bridge between them is the learner's synthesis rather than a source claim: D01 says **what must be visible**, L02 says **in what order it must become visible**, and it should be flagged as such in any interview use. The linter below operationalizes the checklist mechanically; its value is as an argument about what a write-up is for, not as a quality metric.

The non-trivial connection to the approval bottleneck: a write-up that pre-commits to what would change the author's mind is also the write-up that a reviewer can clear in ten minutes, because the reviewer's question is not is this right? but **does the author's stated basis justify the decision they are asking for, and is the residual risk one we can name?** Reasoning transparency is therefore not only an ethical norm; it is the input that makes a short approval cycle possible. That is the mechanism that connects D01 to P01's shortened chain, and it is not stated in either source.

**Step 2**

P01 reports the current shape of the chain: **a typical grant now needs active sign-off from only one person in addition to the primary grant investigator**; decisions were moved closer to the grantmakers with the most context; and steps were cut **especially for grants clearly above our bar**. The comparator cases: ten million dollars approved twelve days after the first conversation, and a technical-AI-safety RFP whose approval process was redesigned before the team tripled its giving from 46 to 157 million dollars across more than 200 grants (SOURCES P01, P06).

The non-trivial reading: the fast path is conditioned on **conviction**, not on grant size. The twelve-day grant is large; so is the 160-million-dollar Resolution grant. A workflow that spends more diligence on small uncertain grants and less on large confident ones is coherent only if the conviction is well calibrated. P06 supplies the calibration claim rather than the proof: the CEO reports that people with short AI timelines have been **more accurate than mine** over recent years and that domain experts have outperformed superforecasters on rate of progress. That is a track-record argument, and it carries an acknowledged risk that P06 names in its own footnote about extrapolative expectations driving bubbles.

There is an apparent tension with D04, which states that on first funding the funder **typically (but not always) prefer[s] to begin with a grant of 200k to 2M per year over one to two years, and then consider renewing**. Large and fast versus staged and small. The reconciliation, which is the learner's inference and should be presented as such, is that these are two positions on one axis: how much is already known. Staging buys information cheaply when quality is uncertain; the fast path is used where the prior is already strong, which in the public record means an existing grantee with a track record (Redwood at more than 70 million dollars; Epoch at 68 million), a founder with a world-class reputation (Resolution), or a case the author is confident enough to sign. Staging is also option value: D04 notes that a top-up can expand a project before the initial period expires if early evidence of success arrives, which is exactly the payoff structure of a tranche.

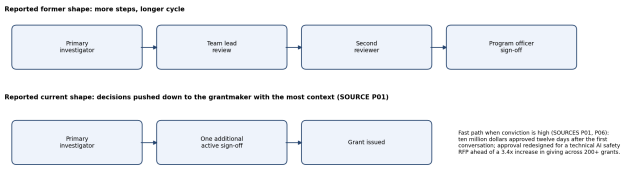


Figure 5: Graphical analysis for Step 2.

**Step 3**

The named binding constraint is grantmaker time. P01 states it directly: the team **remain[s] capacity-constrained and the main thing preventing us from scaling our impact is staff time**. P06 adds that the organisation wants to keep a light staffing ratio while taking on more donors, that it aims to bring on roughly seventy staff, and that it competes for talent against AI labs whose total compensation it cannot match. P06 also names the internal tension created by scarcity: expanding existing grantees and creating new highly absorptive organisations **both compete in the near term for scarce grantmaker time**, and helping create new grantees is **far more labor intensive per dollar moved at the beginning**.

Other constraints stack behind it. Review capacity at volume (D04: raise the bar rather than close the round). Trust: P12 argues that trust will be **the bottleneck to deploy this wave of funding**, notes that the community's mutual trust probably won't scale, and describes the principal-agent problem it calls the cursed game of philanthropy. Absorptive capacity of the field: P10 states flatly that **what's scarce is organizations**. And one constraint that no source lists as a constraint: the time the process imposes on applicants, which P04 and D04 both recognize as a cost but which appears in no dashboard.

Table 14: Bottlenecks and responses as stated in the corpus. The residual-risk column is the learner's inference and should be labelled as such in interview.

| **Bottleneck** | **Evidence** | **Observed response** | **Residual risk (inference)** |
|---|---|---|---|
| Grantmaker hours | P01: staff time is the main constraint; P06: roughly seventy hires planned | Shorter investigations, sign-off pushed down, internal AI tooling (D02) | Reviewer diligence shrinks faster than the work it is meant to check |
| Review capacity at volume | D04: EOIs overwhelmed staff capacity; P02: nearly 500 applications against a forecast of about 40 | Raise the invitation bar; no feedback on most EOIs | Applicants absorb the cost; the funder's field map ages if the funnel narrows |
| Trust between funders and founders | P12: trust as the deployment bottleneck; the cursed game of philanthropy | Regranting, colocating money with taste, transparency by default | Not fully within the funder's control |

| **Bottleneck** | **Evidence** | **Observed response** | **Residual risk (inference)** |
|---|---|---|---|
| Absorptive capacity of grantees | P06: existing grantees and new high-capacity orgs compete for the same grantmaker time; P10: what is scarce is organizations | Proactive seeding, founder calls, incubators, staged grants with top-ups | Origination costs more per dollar moved early, so it pressures the same scarce resource |
| Talent relative to frontier labs | P06: candidates treat lab safety roles as comparably impactful at far higher pay | Pay changes under consideration; support for the wider ecosystem to do the same | Compression of the applicant pool at the judgment-heavy margin |
| Applicant and counterparty time | P04: avoiding unnecessary work for the organizations you are evaluating; D04: count costs the project imposes on others | EOI-first process; fifteen-minute expert calls | Unmeasured, therefore systematically under-counted, and it falls on people with the least slack |

1. Triage: does this deserve evidence-seeking at all, and how much? (SOURCES P04, D04)
2. Evidence selection: which single piece of information would most likely flip the answer, and is it worth its cost to both parties? (SOURCE P04)
3. Bar decision: does the opportunity clear the bar, on the author's stated basis? (SOURCES D04, P01)
4. Route decision: proactive seed, staged first grant with renewal, or large fast grant, and which sign-off path applies? (SOURCES D04, P01, P06)
5. Stop decision: what rule ended the investigation, and would a reviewer agree it was applied consistently? (proposed exercise)

The non-trivial point about the throughput metrics. P06 takes pride in an unusual combination: moving more money and making a large number of grants per staff member. Those are throughput measures, and P01 supplies the reason they cannot be the objective: **our ultimate goal isn't to move money**, and **the dollar figures are a proxy**. Shortening investigations raises throughput

arithmetically, but whether it raises impact depends on how decision quality behaves as time falls. Under a saturating quality curve there is an interior optimum, and cutting further becomes destructive. No source in the corpus estimates that curve, so the honest position is that the trade-off is real, currently unmeasured, and worth a candidate flagging explicitly rather than assuming that faster is better.

$$\text{impact}(h) = \frac{U}{h}\left(1 - e^{-(h/h_0)^2}\right), \qquad U = H_{\text{year}}\,(1 - \text{overhead})$$

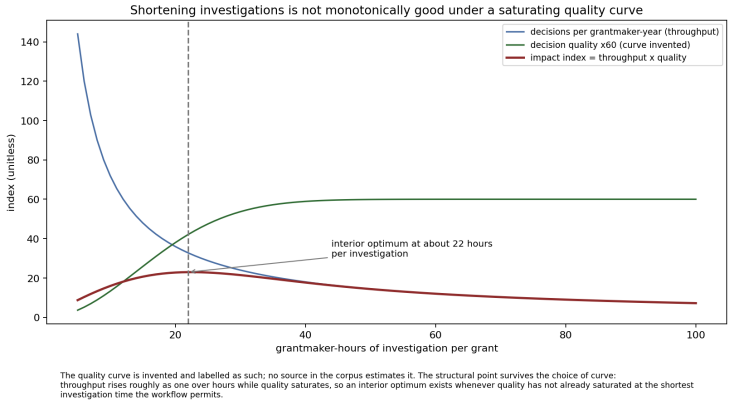


Figure 6: Graphical analysis for Step 3.

### Step 4

D02 describes three internal tools. Grant Copilot gathers information about a grant, generates draft writeups, and preps a grantmaker for their team lead's review. A BOTEC generator turns a grant idea into a back-of-the-envelope cost-effectiveness calculation. Subagent research workflows produce detailed research reports, with one biosecurity researcher reporting a five-to-tenfold speedup. The stated division of labour is explicit: the tools let **the grantmaker focus on important decision-making, rather than information collection**. P06 frames the same programme as a way to keep the staffing ratio light, citing Claude skills for grant writeups and cost-effectiveness calculations.

D02 also supplies the caveats, and they are unusually candid. The tools are **still quite poor at writing serviceable prose**, with an uncanny, sales-driven quality; the author tried to draft the very post in question with AI and started over. Models still get confused or lack context. And: **AI can also make it feel like you're working more quickly, when you're actually just writing more text without accomplishing anything useful**. The closing position

is that the organisation still needs **to use our own writing as a tool to think deeply**.

The non-trivial consequence for the workflow. If information collection is automated and judgment is not, then the marginal value of a grantmaker-hour shifts almost entirely onto judgment, which raises rather than lowers the bar on exactly the competencies the brief names. The interesting interview answer is not I would use the tools but a specification of what must not be automated and why: the write-up, because writing is how the author finds out whether they actually believe the recommendation, and the counterparty conversation, because it produces trust and contextual read alongside information. Both of those judgments are the learner's inference from D02's stated caveats rather than claims the source makes.

## Sub-Question 4: After the Yes: Relationship Management, the Loop Back, and a Proposed Stop Rule

The process does not end at approval. The corpus describes success criteria agreed with grantees, encouragement to raise ambitions, renewals and top-ups, and an unusual worry about which parts of the ecosystem the funder's own feedback loops select for.

Reconstruct the post-approval work, examine the loop it creates back into deal flow, and then convert the whole reconstruction into the promised exercise: a decision-relevant investigation checklist and a defensible stop rule, with a competitor.

### Step 1

Levin's representative Tuesday contains a call with a grantee **to align on success criteria for their project**. Chaver describes a collaborative relationship and an aim of amplifying impact by identifying the right goal or bringing in key talent (SOURCE P04). P01 describes the same posture at fund scale: supporting existing grantees as they raise their ambitions, encouraging them to think about how they could spend more, including larger compute budgets and higher salaries to compete for talent, with the explicit qualification that **we don't think every organization or idea should scale**.

The non-trivial point is that **renewal is a different decision from origination**, with a different evidence set and a different failure mode. The evidence shifts from the plausibility of a plan to observed execution, the marginal use of the next dollar, absorptive capacity, and the talent market in which the grantee is hiring. The failure mode shifts in magnitude: a mistaken first grant is bounded, while a mistaken renewal at seventy million dollars is not. Chaver's frame, that a grant should be judged by whether it clears the bar, is therefore incomplete for renewals: at that stage the question is whether the grantee's marginal dollar is still the funder's best marginal dollar, which is a portfolio

question, not a project question. P06 gives the vocabulary for this: the historically marginalist mode of equalising returns across funds, and the shift toward **more ambitious goals** and inframarginal grants where the funder had to recruit a founder or incubate an organisation before anything fundable existed.

**Step 2**

Successful grantees generate new opportunities: renewals, top-ups, referrals, and spin-outs. P05 describes an explicit broadening of how impact is counted for exactly this reason: beyond direct return on investment, the ecosystem can contribute **by influencing major funders, strengthening policy relationships, incubating organizations that later secure large-scale funding, and building a talent pipeline**.

The non-trivial consequence is a **selection effect**: the pipeline is fed disproportionately by organisations the funder already knows, whose outputs are already measurable, and whose leaders are already trusted. That is efficient, and it is also a bias. P02 supplies the evidence that the bias is real, because the strongest effective-careers opportunities came from **impact-focused organizations well outside the EA network, several of which were not previously on our radar**. P05 names the corresponding worry in its own open questions: ***Are we appropriately balancing the shorter feedback loops of scaling donation flows against the longer-term, harder-to-measure impacts of shaping careers, or are we being drawn too much toward what's easier to evaluate?*** The open call is the corrective instrument, and its measured value is precisely that it purchases the funder's own forecast error.

S10's list of philanthropic pitfalls belongs at this stage rather than at the front of the process: fund based on cost rather than value delivered; a game of chicken between funders producing glacial response times; unclear credit allocation and double counting of impact. Every one of these is a consequence of many funders managing overlapping relationships with the same small set of grantees, which is exactly what a dense, high-trust ecosystem produces when it scales. P12 predicts that this density will **probably not scale** with the money.

**Step 3**

This is the proposed exercise. No internal investigation template is in the corpus, so the checklist below is a reconstruction assembled from D01's transparency requirements, Levin's evidence-selection rule, D04's assessment criteria and D02's tooling description. It should be presented in interview as a proposal, not as a description of how the organisation works.

Table 15: Reconstructed as a proposed exercise. Items one, two, three and eight derive from D01 and P04; item five derives from D04; the remainder are the learner's synthesis.

| **Checklist item** | **Question it answers** | **Failure mode it prevents** |
|---|---|---|
| Name the decision and the decision-maker | Who signs, and what exactly are they being asked to approve? | A memo that reports on the world instead of asking for a decision |
| State the decisive consideration and the nearest competitor | If only one reason survived, which would it be? | Buried reasoning; the reviewer cannot see what actually carried the weight |
| Pre-commit to the evidence that would change your mind | What would have to be true for a no to become a yes? | Confirmation-focused investigation |
| Give a confidence level and a support kind per major claim | Was this a skim, an expert call, a controlled comparison or an intuition? | Confidence laundering, where a guess reads like a finding |
| Count the hours imposed on the applicant | What did this investigation cost the counterparty? | Over-investigation and applicant fatigue, which D04 treats as a project cost |
| Name the strongest counter-case you have funded elsewhere | Would we fund the opposing view too, and have we? | Silent selection into a single worldview, contrary to P08's stated pluralism |
| Say what follows on a yes and on a no | Renewal, tranche, referral to another funder, or decline? | Grants issued without success criteria or an exit condition |
| State the stop rule that ended the investigation | Why stop here rather than earlier or later? | Unbounded investigation and unauditable effort allocation |

Two defensible stop rules, with the objection that each must survive.

- Time-boxed. Pre-commit a number of grantmaker-hours to each grant by size or by stage, and escalate only on a named trigger. Defence: it is legible, auditable, protects throughput and matches the observed practice of shortening investigations and pushing decisions down (SOURCE P01). Objection: time is a proxy for the value of information, and it fails on precisely the case that matters, where one more hour would flip a large decision. It also treats all hours as equal, ignoring that an hour of the ap-

plicant's time may be worth more than an hour of the funder's (SOURCE P04).

- Evidence-sufficiency. Stop when the expected value of the next evidence-gathering action falls below its cost, counting both parties' time. Defence: it targets the decision rather than the clock, and it is the natural formalization of Levin's rule (SOURCE P04). Objection: it requires a prior, a likelihood and a value of information, all of which are noisy, and it is self-serving in a specific way, since an investigator can always declare the next piece of evidence low-value. It also degenerates into confirmation-seeking if the questions posed are only the ones whose answers the investigator can already predict.

The synthesis worth stating aloud: use the time box as the default and the value-of-information test as the escalation trigger. The time box protects against unbounded investigation and against the asymmetry in which the applicant's hours go uncounted; the escalation test preserves the ability to spend more where the decision is large and the evidence is genuinely decisive. Each rule covers the other's weakness, and naming which weakness each one covers is a stronger answer than defending either alone. The exercise is therefore not to choose a stop rule but to specify the trigger that moves an investigation from the default regime into the escalation regime, and to make that trigger reviewable after the fact.

# Map CG's donor, advisory and partnership context relevant to evaluator work.

## Question

What map of Good Ventures, external philanthropists, the advisory and partnerships function, co-funders, and fund intermediaries is needed for evaluator and grantmaker work at Coefficient Giving, and what implications and conflicts arise for grant recommendations?

## Background

This subquestion maps the donor-side and advisory-side context that a grant evaluator must model. The central source anchors are P06, the 2025 CEO letter; S07, the Managing Director, Philanthropic Advisory hiring post; D04, the AI Governance request for proposals; P10, the EA Funds Transformative AI Fund launch; P11 and P12, Manifund regranting and the money-taste-dealflow-hustle-trust frame; P07, individual staff giving suggestions; and S01, S03, S10, N02, P13, and D03 for co-funding and intermediary examples.

The non-trivial principle is that Coefficient Giving is a **recommender and intermediary**, *not a single unitary donor. An evaluator is matching opportunities to multiple principals with different constraints, time horizons, and appetites. Source figures and role descriptions are historical and may not reflect current vacancies, fund boundaries, or donor preferences.*

## Sub-Question 1: Stakeholder Map: Money, Advice, Trust, and Conflicts

Build a stakeholder map of Good Ventures, external philanthropists, the advisory and partnerships function, co-funders, and fund intermediaries or regrantors. For each, state the role, source anchor, implication for a grant recommendation, and characteristic conflict or caveat.

Then add a conflict register and a rehearsal exercise. Distinguish what the sources state from what is an inference about evaluator behavior. Use P06 and S07 for institutional design, D04 for recommendation-sharing mechanics, and P10, P11, P12, P13, and D03 for intermediary and regranting dynamics.

### Step 1

**1. Anchor the money map: CG is a multidonor recommender, not a single endowment.**

P06 states that Coefficient Giving directed over $1 billion in 2025, that Good Ventures is its largest funding partner, and that funding from non-Good Ventures donors more than doubled to over $200 million. D04 says a strong application might be funded by Good Ventures or by any of more than 20 other

philanthropists who have asked to hear about grant opportunities. The non-trivial jump is that the operative question is not simply whether CG has money, but which principal can fund a given opportunity, under what constraints, and with what sharing permission.

Table 16: Stakeholder map for donor, advisory, and partnership context. Source anchors are historical public documents.

| **Stakeholder or channel** | **What it is** | **Source anchor** | **Implication for a recommendation** | **Characteristic conflict or caveat** |
|---|---|---|---|---|
| Good Ventures | Largest funding partner; can fund CG-recommended grants | P06; D04 | A recommendation may become a Good Ventures grant; its capacity and preferences shape feasibility | Donor intent versus CG cost-effectiveness bar; committed versus conditionally approved versus published timing |
| More than 20 external philanthropists | Receive recommendations only if the applicant permits sharing; give far less each year than CG; interested in narrow subsets; in aggregate do not act on most recommendations | D04 | Sharing permission materially changes expected funding odds but does not guarantee funding | Confidentiality; applicant expectations; fragmented portfolio; no single donor sees the whole field |

| Stakeholder or channel | What it is | Source anchor | Implication for a recommendation | Characteristic conflict or caveat |
|---|---|---|---|---|
| Partnerships and Philanthropic Advisory | Partnerships team grew from 2 to 10; S07 seeks a Managing Director to build a bespoke advisory practice from roughly scratch, lead a 12-person team, and sit on the leadership team | P06; S07 | Evaluator work feeds a donor-facing advisory function, not only an internal allocation process | Advisory service versus independent evaluation; general-manager ambition versus marginalist discipline |
| Co-funders | Other funders who join all or part of a grant, such as other GiveWell donors, SFF in a matched campaign, Longview, or BlueDot | S01; N02; S03 | Co-funding can leverage CG dollars and provide a second opinion, but the recommendation must remain clear about who is funding what | Credit allocation; double-counting of impact; game-of-chicken delays between funders |

| Stakeholder or channel | What it is | Source anchor | Implication for a recommendation | Characteristic conflict or caveat |
|---|---|---|---|---|
| Fund intermediaries and regrantors | EA Funds Transformative AI Fund, Manifund regranting, BlueDot Rapid Grants, and similar small or early-stage channels | P10; P11; P12; P13; D03 | They absorb early-stage and small grants that are often below CG comparative advantage, and can be a referral path | Boundary confusion; duplication; attribution and counterfactual difficulty |
| Individual program staff suggestions | Holiday giving suggestions from individual staff or teams | P07 | Useful signal of staff judgment, not an institutional funding recommendation | Individual view may be read as CG endorsement |

**Step 2**

**2. External philanthropist channel: sharing increases odds but imports donor-specific constraints.**

D04 states that applications are substantially more likely to be funded if applicants give permission to share them with other funders. But the same passage says those philanthropists give far less each year than CG, are typically interested in only a narrow subset of opportunities, and in aggregate do not act on most recommendations. It also says the area remained funding constrained at the time of writing, with total philanthropy motivated by AI catastrophic-risk mitigation probably under $200 million per year.

The non-trivial interpretation is that sharing permission is a design variable in the recommendation, not an administrative afterthought. It changes expected value, creates confidentiality duties, and may require tailoring the write-up for different donor screens without changing the underlying analysis.

- Ask about sharing permission early, because a no-sharing answer can foreclose the main external funding path.

- Do not infer donor interest from CG interest. CG may like a proposal that no external philanthropist will fund.
- A recommendation to an external funder can create adverse-selection issues if only rejected or difficult cases are shared.
- Maintain reasoning transparency so external funders can assess the recommendation without redoing the full investigation.
- Distinguish donor-specific constraints from field-level funding constraints when deciding whether to decline or reroute a proposal.

**Step 3**

**3. Advisory and partnerships function: from marginal funder to general manager.**

P06 says the partnerships team grew from 2 to 10, and S07 says CG is hiring a Managing Director, Philanthropic Advisory to build a bespoke advisory practice from roughly scratch, lead what is described as a 12-person partnerships team, and sit on the leadership team. S07 explicitly says this is not a conventional fundraising background; it wants someone with the capability and drive to be a general manager of directing new funding to impactful causes.

The non-trivial jump is that this function changes the evaluator's operating environment. An evaluator may be producing material for external advisory relationships, not only for internal allocation. That makes legibility, conflict disclosure, and donor-specific constraint handling part of the craft.

P06 also warns against the strategic funder trap: thinking CG has all the answers and slotting grantees into a vision almost like subcontractors. The CEO letter presents the marginalist approach as relatively strategy-agnostic and open to grantees' local context. The advisory function can pull in the opposite direction, toward bespoke donor service and funder-designed initiatives.

- Expect to co-develop proposals and relationships, not only score finished applications.
- Treat donor management as a distinct function that consumes evaluator output; write so that another person can use the reasoning.
- Watch the tension between bespoke donor advice and cause-level prioritization across funds.
- Preserve reasoning transparency and grantee local knowledge even when the output is donor-specific and not publishable.

**Step 4**

**4. Co-funders, fund intermediaries, and regrantors: complementary infrastructure with attribution costs.**

P10 says the EA Funds Transformative AI Fund will primarily fund early-stage projects, with grants typically below $150,000, rare grants above $300,000, and no grants above roughly $1 million for the foreseeable future. P11 says Manifund regranting has made first checks to organizations that later grew, including Timaeus, which later merged with Resolution and received a $160 million CG grant. P12 frames funding as needing money, taste, dealflow, hustle, and trust, and says trust is a likely bottleneck. D03 says BlueDot Rapid Grants go up to $10,000 with fast decisions.

The non-trivial interpretation is that these intermediaries are not simply competitors for dealflow. They are part of the funnel that produces the organizations CG may later fund at larger scale. That creates a real counterfactual question: if CG funds the later stage, how much credit belongs to the earlier small funder, and how should that affect impact accounting?

Table 17: Intermediaries and co-funders as funnel infrastructure, not only as alternative funders.

| **Intermediary or co-funder** | **Typical role and scale** | **Source anchor** | **Evaluator implication** | **Conflict or caveat** |
|---|---|---|---|---|
| EA Funds Transformative AI Fund | Early-stage, fast, failure-tolerant grants typically below $150k; rare above $300k; no above roughly $1m | P10 | Complement to CG's larger grants; useful referral path for very early projects | Boundary between funders can blur; absorptive capacity may limit scale |
| Manifund regranting | Part-time regrantors making many small, fast, hits-based grants | P11; P12 | Can surface founders and projects before centralized funders see them | Credit and double-counting; principal-agent trust; 'cursed game of philanthropy' |

| Intermediary or co-funder | Typical role and scale | Source anchor | Evaluator implication | Conflict or caveat |
|---|---|---|---|---|
| BlueDot Rapid Grants | Grants up to $10k; decisions usually within about five working days; field building, compute, events, travel | P13; D03 | Useful for small bottlenecks and talent retention; may reveal future CG grantees | Small grants are generally outside CG comparative advantage; duplication risk |
| Matched or co-funding examples | Other donors join a grant or unlock a match, e.g. GiveWell donors in S01, SFF in N02, Longview and BlueDot in S03 | S01; N02; S03; S10 | Leverage and validation; can make a marginal grant viable | Credit allocation; funder game-of-chicken; unclear marginal attribution |

**Step 5**

**5. Conflict register and rehearsal exercise.**

The map is only useful if it changes evaluator behavior. The following register names the recurring conflicts and the mitigation an evaluator can practice. The mitigations are inferences from the sources and from general grantmaking practice, not claims that CG has already adopted each one.

Table 18: Conflict register for donor, advisory, and intermediary relationships.

| **Conflict** | **Why it arises** | **Source anchor** | **Mitigation to rehearse** |
|---|---|---|---|
| Donor preference versus cost-effectiveness bar | Multiple principals with different criteria receive the same recommendation | D04; P06 | Separate the analysis from the funding recommendation; state donor-specific constraints explicitly |
| Advisory service versus independent evaluation | Partnerships function grows and serves external donors | S07; P06 | Disclose roles and conflicts; keep reasoning transparency even in non-public advice |
| Credit and double-counting | Co-funding and regranting create overlapping causal claims | S10; P11 | Use counterfactual accounting; do not sum impact across funders without adjustment |
| Confidentiality versus transparency | Sharing with external funders exposes applicant material | D04; D01 | Share only with permission; redact sensitive details; publish reasoning when possible |
| Principal-agent trust | Money and taste or dealflow can sit with different actors | P12 | Build track records; share dealflow; consider colocating money with taste where appropriate |

| **Conflict** | **Why it arises** | **Source anchor** | **Mitigation to rehearse** |
|---|---|---|---|
| Committed versus conditionally approved versus published | Grant accounting and public databases lag and use different definitions | P01 footnote 2 | Use precise language; do not treat all announced dollars as identical or final |

Exercise. Take a hypothetical early-stage AI safety proposal. First classify each constraint as donor-specific, field-level funding constraint, institutional process, or legal and confidentiality. Second, write a 150-word sharing recommendation that states what may be shared, with which class of funder, what must be redacted, and what you need from the other funder in return. Third, give two defensible answers to the question: should this be recommended to CG for a larger grant, or routed to a fund intermediary? A direct CG recommendation answer rests on scale, absorptive capacity, and CG comparative advantage. A route-to-intermediary answer rests on stage, small budget, fast decision needs, and preserving CG grantmaker time.

Calibrated answer script. I would map the funding principals first, because CG is a recommender across Good Ventures and external philanthropists rather than a single pool. I would ask about sharing permission early, note that it raises expected funding odds but does not guarantee funding, and separate the cost-effectiveness analysis from donor-specific constraints. I would flag conflicts around advisory service, credit allocation, confidentiality, and principal-agent trust, and I would preserve reasoning transparency wherever the material can be shared. Finally, I would treat fund intermediaries and regrantors as complementary funnel infrastructure, while adjusting impact accounting for the earlier small checks that made the organization possible.

# Extract hiring signals and cultural expectations from the role brief and CG and aligned-actor posts.

## Question

Extract the **hiring signals** and **cultural expectations** that can reasonably be read from the attributed role brief, Coefficient Giving institutional posts, individual staff posts, and aligned-actor posts. Produce a checklist of valued traits with examples and with explicit caveats about how much weight each signal should carry.

The central problem is not compiling adjectives. It is sorting evidence by **source type and role context**, because a trait valued by one team or in one historical hiring post may not transfer to a current evaluator interview.

## Background

The role brief is an attributed historical summary, not the full advertisement. It states requirements such as ownership of ambiguous work, ambition and speed, interpersonal adaptability, familiarity with catastrophic-risk communities, probabilistic expected-value reasoning, decisive consideration identification, proportionate due diligence, and concise calibrated transparent communication.

CG posts add institutional culture: P04 shows grantmaker workflows and norms such as asking what evidence would change one's mind, constructive thought partnership, execution mindset, and speed to decision; P06 stresses recruiting and empowering exceptional people, agility, risk tolerance, general-manager ownership, avoiding the strategic-funder trap, and light staffing; P10 and S02 from EA Funds stress agency and field knowledge; S07 stresses a general manager of directing new funding; S08 stresses resisting risk aversion; S09 shows Project Tailwind's founder-call tiers.

Caveats: the hiring posts are historical, P04's deadline has passed, sources describe different roles and teams, and none establish a current vacancy or an application by the learner. Treat signals as hypotheses about what an evaluator may reward, not as personal credentials.

## Sub-Question 1: Extract Hiring Signals and Cultural Expectations

Task: extract a source-anchored hiring-signal checklist. For each signal, record the source anchor, the trait or behavior it likely rewards, and the caveat that limits its weight. Then identify which signals are robust across the role brief, CG institutional voice, and aligned-actor posts, and which are role-specific or historically bounded.

Non-trivial move: **separate source tiers**. The role brief is an attributed role description; CG institutional posts speak for the institution; individual staff

posts speak for a person or team; aligned-actor posts reveal norms in the broader funding ecosystem. A signal repeated across tiers is stronger than a signal found in one post, but repetition can also reflect a shared subculture rather than an interview rubric.

**Step 1**

The first non-trivial move is to separate source tiers before extraction. Flattening all sources into one voice would overstate the evidence. Use the following tiering.

- Tier A: attributed role brief. Supports stated competencies and role requirements, but not a current vacancy and not personal fit.
- Tier B: CG institutional posts and RFPs. P01, P04, P05, P06, P08, D02, D04, S01, S04, S07, S08, S09. Supports institutional culture, scale, priorities, and stated expectations, but may be outdated and may aggregate multiple teams.
- Tier C: individual staff posts and interviews. P04 staff narratives, S02, S03, S05, S06, S09, S10, S11. Supports concrete behaviors and local norms, but is individual and role-specific.
- Tier D: aligned-actor posts. P10, P11, P12, D03, and N01 through N06 where relevant. Supports ecosystem norms and adjacent expectations, but not CG's own rubric.

Then extract signals only at the tier they can support. For example, **agency** is explicit in P10 and S02 for EA Funds, and is consistent with the role brief's ownership language; that cross-tier consistency raises its weight for evaluator roles, but it does not make it a CG-wide formal criterion.

**Step 2**

The second move is to convert adjectives into observable behaviors. A hiring signal is more useful when it implies a question an interviewer could ask or a behavior they could reward.

| Signal | Source anchor | Likely rewarded behavior | Limits on weight |
|---|---|---|---|
| Ownership of ambiguous work | role brief; P04 Abbey Chaver; P06 general-manager section | defining problem, identifying dependencies, buying in stakeholders, measuring success, acting without being asked | broad; every grantmaker role may weight it differently |
| Agency | P10; S02 | taking initiative without being asked, sourcing projects, building processes | explicit for EA Funds role; transfers as hypothesis not guarantee to CG evaluator |
| Ambition and speed | role brief; P01; P06 | bigger grants, faster investigations, 12-day approval example, scaling | speed is institutional goal; may conflict with due diligence and proportionality |
| Calibrated communication | role brief; D01; D04 | stating confidence, linking claims to support, separating inference from source | D01 is earlier Open Philanthropy and may be outdated; norms persist in D04 criteria |
| Reasoning transparency | D01; D04 | making reasoning legible, identifying decisive considerations, scope sensitivity | one institution's norm; interview may test it but not formally score it |
| Evidence that would change one's mind | P04 Trevor Levin | asking early what evidence would change the mind, seeking that evidence | staff habit; good interview behavior but not a stated role requirement in the brief |

| Signal | Source anchor | Likely rewarded behavior | Limits on weight |
|---|---|---|---|
| Interpersonal adaptability | role brief; P04 | working with grantees, experts, funders, applicants; thought partnership | broad; difficult to infer from a single source |
| Judgment of people and organizations | role brief; P04 Chris Bakerlee reference checks; P06 early bets | assessing founders, track records, references, project leads spread thin | role brief says evaluation requires judging people; evidence of how to do it is illustrative |
| Catastrophic-risk familiarity | role brief; P01; P08; D04 | knowing AI safety, biosecurity, governance, threat models, field actors | domain breadth is huge; familiarity is not expertise in every subfield |
| Proactive dealflow and founder recruitment | P01 Project Tailwind; P04; S09; S05 | finding would-be founders, seeding organizations, active grantmaking | proactive posture is CG strategic choice; may vary by team and year |
| General-manager ownership | P06; S07 | owning a problem end-to-end, building programs, directing funding | P06 warns against strategic-funder trap; S07 is a different role |
| Risk tolerance while scaling | P06; S08 | making weird-seeming bets, resisting risk aversion, agility | stated aspiration; scaling can pull the other way |
| Light staffing and efficiency | P06; D02 | high grants per staff member, AI enablement, fewer approval steps | organizational metric; not a personal hiring rubric |

| Signal | Source anchor | Likely rewarded behavior | Limits on weight |
|---|---|---|---|
| Constructive thought partnership | P04 Abbey Chaver | amplifying grantee impact, identifying right goal, bringing in talent | one grantmaker's practice; likely valued but not universal |
| External representation | S07 | representing CG to major donors, strategic sharpness, executive capability | role-specific to Managing Director, Philanthropic Advisory |
| Speed to decision | P01; P04; D03 | shortened investigations, one additional sign-off, rapid grants | speed can be domain-specific; grants above bar get faster treatment |
| Field knowledge | P10; S02 | knowing AI landscape well enough to have own views | explicit for EA Funds; CG role brief asks familiarity, not omniscience |

The non-trivial interpretation is that the same phrase can be a hiring signal in one source and a strategic objective in another. **Speed** in P01 is an institutional response to AI timelines; in an interview it may show up as a question about proportionate due diligence. **Agency** in P10 is an explicit hiring criterion for a different fund; in the CG role brief it is implied by ownership of ambiguous work. Do not over-weight adjacent-actor language.

**Step 3**

The third move is to turn the table into a rehearsal checklist. Use the checklist to generate evidence from general reasoning and from the supplied material, never from invented personal credentials or work history.

- Show ownership by narrating how you would define a problem, map dependencies, identify stakeholders, and measure success, using the P04 product-manager-to-grantmaker translation as a model.
- Show agency by describing how you would source opportunities rather than wait for applications, consistent with P10's open-application and early-stage funding logic and P01's proactive seeding.

- Show calibrated communication by stating confidence, naming what support you have, and separating source-supported claims from inferences, as D01 and D04 emphasize.
- Show evidence-seeking by asking what piece of evidence would most change your mind and how you would get it efficiently, as P04's Trevor Levin describes.
- Show speed with proportionality by explaining when a grant clearly above bar can move quickly and when due diligence should slow it down.
- Show catastrophic-risk familiarity by using the supplied corpus to discuss AI safety, biosecurity, governance, and evaluation concepts without pretending to possess credentials.
- Show interpersonal adaptability by describing how you would work with disagreeing experts, grantees, and funders, consistent with P08's support for a wide range of experts and P04's thought partnership.
- Show judgment of people and organizations by discussing references, track record, theory of change, and failure modes, as in P04's ERA and founder-transition examples.
- Show risk tolerance while scaling by acknowledging the tension between agility and risk aversion in P06 and S08.
- Show external representation only if relevant to the role; S07's Managing Director, Philanthropic Advisory signal is role-specific.

Caveats that limit weight: P04's grantmaker hiring deadline has passed; P04, S07, P10, and S02 describe different roles and organizations; P06 is a CEO letter, not a rubric; D01 is an earlier Open Philanthropy document; P01 and P05 contain figures that will age; and the role brief is historical. Treat every signal as a hypothesis to test, not as a guaranteed interview criterion.

**Step 4**

The fourth move is to use the checklist under uncertainty. A hiring-signal checklist is not a personality quiz. Its value is to help you allocate preparation time and to make your reasoning transparent when you discuss fit.

- High-confidence cross-tier signals: ownership or agency, calibrated reasoning transparency, evidence-seeking, speed with proportionate due diligence, and catastrophic-risk familiarity appear in the role brief and multiple CG or adjacent sources.
- Medium-confidence institutional signals: proactive seeding, founder recruitment, general-manager ownership, risk tolerance while scaling, and constructive thought partnership appear in CG posts but may be team-specific.

- Low-transfer signals: light staffing ratio, AI enablement tools, external representation to donors, and specific fund structures are role-specific or organizational and should not be treated as personal requirements.
- Disconfirming evidence to watch for: a role may weight deep domain expertise over generalist ownership, or may weight careful due diligence over speed, even though CG posts celebrate both.
- Use in rehearsal: prepare two-minute answer scripts that state a claim, name the source anchor, give the reasoning, identify what would change your mind, and state a limitation. This mirrors D01's reasoning-transparency recommendations without inventing experience.

Final limitation: this step extracts signals from documents. It does not establish that any interviewer uses these signals, that a vacancy exists, or that the learner has the claimed experience. The correct output is a calibrated checklist plus explicit uncertainty about transfer.

# Internalize reasoning-transparency standards and practice applying them to grant memos.

## Question

This step is an internalization drill. You are not being asked to produce a real grant recommendation. You are being asked to make the reasoning-transparency standards in D01 operational, then practice applying them to a grant-memo paragraph.

The deliverable is two artifacts. First, a reasoning-transparency checklist that a grantmaker could run before sending a memo. Second, a rewritten grant-memo paragraph that opens with linked takeaways, flags the most important considerations, states confidence and the kind of support behind major claims, and identifies cruxes.

Success condition: a decision-maker can read the rewritten paragraph, see the recommendation, know how much weight to put on each claim, identify what would change the recommendation, and audit the support without reconstructing the writer's reasoning.

## Background

D01 defines reasoning transparency as making it easy for the reader to answer: How should I update my view in response to this? Its three primary recommendations are to open with a linked summary of key takeaways, indicate which considerations are most important for those takeaways, and indicate confidence in major claims plus the kind of support behind them.

D01 distinguishes confidence expressions. Rough words such as plausible, likely, unlikely, and very likely are cheap but ambiguous across readers. Explicit probabilities or confidence intervals are more precise but cost more and can create false precision. Examples include a 70% confidence interval of 10-120 years for years to high-level machine intelligence, and a 70% confidence statement about the number of systematic reviews the author failed to find.

D01 also distinguishes kinds of support. The support behind a claim might be careful examination of studies the author can assess, shallow skimming of studies the author cannot assess, expert opinion, verifiable facts, a vague impression, a general intuition, a simple robust argument, or a source remembered but not checked. The point is not to rank support mechanically; it is to let the reader see what kind of thing is doing the work.

D04 makes reasoning transparency a grant criterion under strategic judgment, alongside scope sensitivity. P04 shows the practice as a habit: ask early and often what piece of evidence would most likely change your mind, then seek it efficiently.

L02 supplies the writing-craft layer. A memo should give the reader affordances, disambiguate where meanings compete, keep an action close to the expression that prepares for it, give grammatical agency to the decision-maker, and follow a six-move argument structure: meet the reader where they are, take them to the claim, show what could move you, show the evidence, show whether the evidence moves you, and show where the reader should now be.

Caveats: D01 was published under the former name Open Philanthropy and some content may be outdated; CG's internal memo templates are not supplied; and the rewritten paragraph below is a constructed exercise, not a real recommendation. The exercise uses N03 and P13 as reference classes only.

## Sub-Question 1: A reasoning-transparency checklist for grant memos

Construct a checklist that translates D01's primary and secondary recommendations into a pre-send review for a grant memo. The checklist should be usable at three levels: a 60-second scan, a 10-minute self-review, and a full audit for a large or controversial grant. It should force the writer to separate takeaway, importance ranking, confidence, kind of support, and crux.

### Step 1

D01's three primary recommendations are not formatting advice. They are an architecture for the reader's inferential path.

- Open with a linked summary of key takeaways. In a memo, this is not an executive summary of topics. It is a claim-bearing abstract: each takeaway is a sentence that states what the memo recommends or concludes, and each takeaway is linked to the section that supports it.
- Indicate which considerations are most important. The memo should say which two or three claims carry the recommendation. If everything is highlighted, the decision-maker has to reconstruct the priority order.
- Indicate confidence in major claims and the kind of support behind them. Confidence without support is a mood; support without confidence is an evidence dump.

Non-trivial jump: the linked summary is a commitment device. Once the writer lists the takeaways that carry the recommendation, later sections must either support those takeaways or revise them. This makes it harder to bury a weak claim in a long diligence narrative.

Apply D01's own example: the GiveWell Against Malaria Foundation review opens with linked key points, provides detailed responses to major questions, includes a research-process section, and has 125 endnotes. That is an extreme model. D01 explicitly says most analysts should not pay that cost. The checklist should therefore include a proportionality test, not a maximization rule.

**Step 2**

Table 19: D01 confidence ladder adapted for grant memos.

| **Confidence expression** | **Rough meaning** | **When it is enough** | **Failure mode** |
|---|---|---|---|
| Plausible | Non-trivial but not >50%; reader should not rely on it | Peripheral claim; low cost to be wrong | Used to smuggle a central claim without commitment |
| Seems likely | >50% but not investigated in detail | Common-sense or widely believed claim not central to decision | Reader over-weights because it sounds like a finding |
| Likely or very likely | High confidence; usually should be paired with support | Major claim with assessable evidence | Confidence adjective does the work that evidence should do |
| 70% confidence interval | Explicit range; communicates both point estimate and spread | Major quantitative claim where spread matters | False precision if the model is unstable |
| At least 10% with moderate robustness; at least 1% with high robustness | Two confidence levels for two robustness assumptions | Load-bearing premise where robustness is disputed | Complexity can obscure the headline probability |
| I have not checked this | Explicit non-support | Claim is not load-bearing or is cheap to verify later | Used too often until the memo has no load-bearing claims |

D01's examples show the ladder in use: a 70% confidence interval of 10-120 years for years to high-level machine intelligence; a 70% confidence statement about the number of systematic reviews the author failed to find; a seems likely for the importance of sleep; and a plausibly for CBT-I being the most common behavioral treatment for insomnia in clinical practice.

Table 20: D01 support taxonomy reduced to memo categories.

| **Kind of support** | **What the reader learns** | **Memo use** |
|---|---|---|
| Careful examination of studies you are qualified to assess | Direct evidence; author can evaluate methods | Carries major claims |
| Careful examination of studies you are only weakly able to assess | Direct evidence but bounded author competence | Major claim with named uncertainty |
| Shallow skimming of studies | Low-depth evidence | Background or scoping claim |
| Expert opinion you can assess | Judgment from a relevant expert | Useful when primary evidence is sparse; name the expert |
| Expert opinion you cannot easily assess | Borrowed judgment | Use only if the expert's track record is the support |
| Verifiable fact with source | Reader can check | Best for factual building blocks |
| Vague impression or general intuition | Author's prior | State it as such; never as evidence |
| Simple or complex argument that seems robust or questionable | Reasoning quality | Use for structuring; flag disagreement |
| Source remembered but not checked | Pointer for verification | Say so; do not let it carry a major claim |

Non-trivial jump: confidence and support are two axes, not one. A claim can be high-confidence because it is obvious and weakly supported by evidence, or low-confidence because the evidence is strong but ambiguous. A memo that collapses the two axes makes it impossible for the reader to know whether to challenge the evidence or the interpretation.

Practice rule: for every major claim, write a confidence phrase and a support phrase. If either is missing, the claim is not yet decision-ready.

**Step 3**

D01 asks what could change the author's mind about key takeaways. D04 asks grant applicants to identify failure modes and downside risks, and to prioritize rather than list every possible risk. P04 shows the habit in practice: ask early what piece of evidence would most likely change your mind, then seek it out efficiently.

1. Name the decision. Is the memo recommending fund, decline, fund with

conditions, or investigate further?

2. List the load-bearing claims. These are the claims that, if false, would change the decision.
3. For each load-bearing claim, ask what evidence would change it. Distinguish evidence that would falsify the claim from evidence that would merely reduce confidence.
4. Rank cruxes by expected value of information. A crux is not just uncertain; it is uncertain and decision-relevant and resolvable at reasonable cost.
5. State the update rule. If X, I would recommend fund; if Y, I would decline; if Z, I would fund a smaller pilot.

Non-trivial jump: a crux is not the same as a risk. A risk is a possible bad outcome. A crux is a possible fact that would change the recommendation. A memo can list many risks and still hide the one fact that actually decides the grant. The crux module forces the writer to say which fact that is.

Tie to L02: the crux module is the show the reader how you could be moved move. It gives the decision-maker a criterion for judging the evidence that follows, and it prevents an unfalsifiable recommendation.

**Step 4**

Table 21: D01 secondary recommendations as costed options.

| **Secondary practice** | **What it buys** | **Cost** | **When it is worth it** |
|---|---|---|---|
| Direct quotes and page numbers | Reader can see the exact support without retrieving the source | Low to moderate | Major claims; disputed evidence; sources behind paywalls |
| Data and code | Independent replication and audit | Moderate to high | Quantitative claims carrying the recommendation |
| Archived copies of sources | Link rot protection; provenance | Low | External links in published memos |
| Transcripts or summaries of conversations | Makes expert input auditable | Moderate to high; anonymity constraints | When expert judgment carries a major claim |

Non-trivial jump: the secondary practices are not a quality checklist to max-

imize. They are options with costs. D01's GiveWell example is explicitly described as probably more costly than it is worth for most analysts. The right question is marginal: what does this transparency buy per grantmaker-hour, and does the grant size and reversibility justify it?

Link to D04: due diligence is proportionate to grant scale and project risk. A $10,000 rapid grant and a $100 million scale grant should not have the same transparency burden. The checklist should include a size and reversibility test before requiring full secondary practices.

**Step 5**

- 60-second scan: Does the memo open with a linked summary of key takeaways, each stated as a claim rather than a topic?
- 60-second scan: Does the summary identify the two or three most important considerations that carry the recommendation?
- 60-second scan: For each major claim, is there a confidence expression?
- 60-second scan: For each major claim, is there a kind-of-support phrase?
- 60-second scan: Is the central crux stated, with an update rule?
- 10-minute self-review: Can a reader tell what would change the recommendation?
- 10-minute self-review: Are confidence and support separated, or has one been used as a substitute for the other?
- 10-minute self-review: Are the most important considerations ranked, or is everything presented as equally important?
- 10-minute self-review: Does the memo distinguish source-supported claims from the writer's inferences?
- 10-minute self-review: Does the memo state its limitations and open questions?
- Full audit for large or controversial grants: Are major quantitative claims backed by data or code, or by a clear pointer to them?
- Full audit: Are quotes accompanied by page numbers or equivalent locators?
- Full audit: Are external sources archived or otherwise stable?
- Full audit: If expert conversations carry major claims, is there a transcript or summary, or an explicit explanation of why not?
- Full audit: Is the level of transparency proportionate to grant size, reversibility, and decision stakes?

- Full audit: Does the memo avoid the strategic-funder trap of substituting the funder's vision for the grantee's local knowledge?

Use the checklist as an editing instrument, not a scoring rubric. The goal is to make the decision-maker's next inferential step available, not to maximize the number of transparency features.

## Sub-Question 2: Rewritten grant-memo paragraph: before, after, and the decision difference

This is a constructed exercise. It is not a real grant recommendation and does not represent Coefficient Giving. It uses N03's AI Safety Poland update as a reference class for a country-level AI safety fieldbuilding organization, and P13 and D04 for scale and due-diligence proportionality.

Task: take a vague grant-memo paragraph, diagnose its reasoning-transparency failures, and rewrite it so that a decision-maker can see the recommendation, the most important considerations, confidence and support for major claims, and the cruxes.

### Step 1

Original constructed paragraph: We should fund this group. They run AI safety meetups in multiple cities and have built a Slack community. The team is strong and well-connected. They have received a small rapid grant already, which shows they can execute. This seems like a good use of money and will help grow the field in a neglected country. I recommend a grant of $150,000 over one year.

- Buried takeaway: the recommendation appears at the end, and the summary is not linked to support.
- Unranked considerations: team, traction, existing grant, neglected country, and field growth are all presented as if equally important.
- No confidence: seems like a good use of money does not say how likely, how robust, or how much would change if evidence were worse.
- No kind of support: team is strong could rest on interviews, references, a track record, or a general impression.
- No crux: it does not say what evidence would change the recommendation.
- Counterfactuality is implied but not argued: they can execute does not separate the rapid grant from what would have happened otherwise.
- No proportionality: it does not explain why $150,000 is appropriate relative to the reference class and the organization's stage.
- No limitations: it does not say what the writer does not know.

**Step 2**

Rewritten opening: Recommendation: fund this country-level AI safety field-building organization at $150,000 over one year, with a midpoint review. The grant is a scale-up bet on a team that has demonstrated local execution, conditional on two checks.

- Takeaway 1: the team has shown above-reference-class execution on a small rapid grant, converting a small budget into a multi-city event series and an active community; see the traction section.
- Takeaway 2: the bottleneck is coordination time, not demand; the organization has more interested participants than it can serve with volunteer capacity; see the capacity section.
- Takeaway 3: the country is neglected relative to its technical talent pool and EU policy proximity; see the neglectedness section.
- Most important consideration: the reference checks on the executive director and the evidence that the rapid grant caused, rather than coincided with, the expansion.
- Second most important consideration: whether $150,000 is the right scale, given that D04's default first-time range is $200,000-$2,000,000 per year for larger grants, and P13's rapid grants cap at $10,000.

Non-trivial jump: the linked summary separates the recommendation from the reasons. The decision-maker can read the three takeaways and the two key considerations, then decide how deep to go. This is the memo-level equivalent of D01's linked summary of key takeaways.

The most important consideration ranking is itself a claim. If the writer gets the ranking wrong, the decision-maker can challenge the ranking rather than reconstruct the whole memo.

**Step 3**

Table 22: Constructed claim table for the rewritten memo paragraph.

| **Major claim** | **Confidence** | **Kind of support** | **What would change it** |
|---|---|---|---|
| The team has above-reference-class execution | High confidence, moderate robustness | Direct review of event outputs, attendance data, and the rapid grant report; two reference calls with people who have worked with the executive director | Reference calls revealing a pattern of over-commitment, or attendance data that double-counts the same participants |
| The rapid grant caused the expansion rather than coincided with it | Moderate confidence | Comparison with the organization's own pre-grant activity; interviews with the team; no independent counterfactual estimate | Evidence that a partner organization or a volunteer was already organizing the same events before the grant |
| Coordination time is the binding constraint | Moderate confidence | Team's own account; event calendar showing canceled or capped events; no time-use study | Evidence that demand is actually saturated, or that the team lacks a pipeline of speakers and participants |
| The country is neglected relative to talent | Low-to-moderate confidence | Comparison of regional fieldbuilding organizations; N03 as a single-country reference; no systematic landscape scan | A mapping exercise showing several well-funded organizations already serving the same talent pool |

| Major claim | Confidence | Kind of support | What would change it |
|---|---|---|---|
| $150,000 is the right scale | Low-to-moderate confidence | Reference-class comparison to D04's first-time range and N03's small grant; stage-based reasoning | A budget showing the marginal dollar would fund administration rather than additional events, or evidence that a smaller grant would unlock the same growth |

D01-style confidence would sound like: I am about 70% confident that the team has above-reference-class execution, and about 55% confident that the rapid grant was a significant counterfactual cause of the expansion; both estimates are unstable and would move with the reference checks.

Non-trivial jump: the table separates the claim, the confidence, the support, and the update condition. The reader can see that the recommendation does not depend equally on all five claims. The first two carry most of the weight; the neglectedness and scale claims are supporting, not decisive.

**Step 4**

1. Crux 1: Do the reference calls support the claim that the executive director can manage a larger budget and a small staff? If not, recommend a smaller capacity-building grant rather than a scale-up.
2. Crux 2: Is there evidence that the rapid grant caused the expansion? If the expansion was already underway or would have happened anyway, the counterfactual value is lower and the grant is less attractive.
3. Crux 3: Is coordination time actually the binding constraint? If the organization cannot absorb more events because of speaker supply or participant demand, the grant should fund a different bottleneck.
4. Crux 4: What is the marginal use of $150,000? If the next dollar funds overhead rather than additional events, a smaller grant or a different structure may be better.

Alternative defensible rewrite A: a shorter, lower-cost memo for a small grant. For a $10,000-$50,000 grant, the memo could open with a one-sentence recommendation, state one confidence phrase and one support phrase for the main claim, and name the single crux. This is defensible under D04's scale criterion and P13's rapid-grant model. It sacrifices auditability for speed.

Alternative defensible rewrite B: a fuller GiveWell-style memo for a multi-million-dollar scale-up. It would add a research-process section, archived sources, a detailed budget model, and transcripts or summaries of expert calls. This is defensible when the grant is large, hard to reverse, and precedent-setting. It is not the right default for every grant.

Non-trivial jump: the two rewrites are not better and worse in the abstract. They are calibrated to different decision burdens. A memo that applies the GiveWell standard to a $10,000 grant wastes grantmaker time; a memo that applies the rapid-grant standard to a $100 million grant hides decision-relevant uncertainty. L02's decision-burden dimension and D04's scale criterion explain why both are defensible.

**Step 5**

Table 23: Decision-maker dimensions affected by the rewrite.

| **Dimension** | **Original paragraph** | **Rewritten paragraph** |
|---|---|---|
| Recommendation visibility | Appears at the end; reader must infer what is being asked | First sentence states fund, amount, duration, and review point |
| Importance ranking | All considerations presented as equal | Two most important considerations named |
| Confidence | None | Explicit confidence phrases for major claims |
| Kind of support | Adjectives such as strong and well-connected | References, event data, grant report, team interviews, comparative reference class |
| Cruxes | None | Four cruxes with update rules |
| Counterfactuality | Implied by they can execute | Separate claim with moderate confidence and a falsifier |
| Proportionality | Not addressed | Scale reasoning tied to D04 and P13 |
| Limitations | Not addressed | Open questions named, including unstable estimates and lack of systematic landscape scan |

The rewrite does not make the grant more likely to be funded. It makes the recommendation easier to evaluate, easier to challenge, and easier to stop if a crux fails. That is the point of reasoning transparency: it improves the quality of the update, not the persuasiveness of the pitch.

## Sub-Question 3: Calibration drill: rewriting vague claims into transparent claims

This drill is the repeatable practice component. The goal is not to memorize a template but to build the habit of pairing every substantive claim with a confidence expression, a kind-of-support phrase, and a crux where the claim is load-bearing.

Instructions: for each vague claim, produce a rewritten claim that a decision-maker could act on. Then score it against the checklist from Subquestion 1.

### Step 1

1. The team is strong.
2. There is a lot of demand for this work.
3. This grant is cost-effective.
4. The field is neglected.
5. The founder is well-connected.
6. This will probably work.

For each claim, decide whether it is load-bearing. If it is not load-bearing, the correct rewrite may be to delete it or mark it as background. If it is load-bearing, the correct rewrite must state confidence, support, and what would change it.

### Step 2

Table 24: Worked calibration answers for the drill.

| **Vague claim** | **Calibrated rewrite** | **Confidence** | **Kind of support** | **Crux** |
|---|---|---|---|---|
| The team is strong | I am about 70% confident the team can execute a 3x budget increase, based on two reference calls and a review of the rapid-grant report; the estimate is unstable because I have not seen them manage full-time staff. | About 70%, unstable | Reference calls; grant report; no direct management observation | Reference calls reveal overcommitment or weak delegation |
| There is a lot of demand for this work | Event attendance grew from X to Y across four cities, and two events hit venue capacity; I have not tested whether the same people attend repeatedly. | High confidence on attendance; low confidence on unique demand | Event calendar and attendance records; no deduplicated participant list | Deduplication shows a small recurring core rather than broad demand |

| **Vague claim** | **Calibrated rewrite** | **Confidence** | **Kind of support** | **Crux** |
|---|---|---|---|---|
| This grant is cost-effective | At $150,000 for an estimated Z additional engaged participants, the cost per participant is within the range of comparable fieldbuilding programs, but the counterfactual is uncertain. | Low-to-moderate | Reference-class comparison; BOTEC; no randomized or quasi-experimental estimate | A credible counterfactual shows most participants would have engaged anyway |
| The field is neglected | I found three comparable country-level organizations in the region, none with more than one full-time staff member; this is a shallow scan, not a systematic landscape review. | Low-to-moderate | Shallow scan of public sources; N03 as one reference | A systematic scan reveals well-funded organizations serving the same talent pool |

| Vague claim | Calibrated rewrite | Confidence | Kind of support | Crux |
|---|---|---|---|---|
| The founder is well-connected | The founder has worked with two named researchers and one policy office, according to public records and one conversation; I have not verified the depth of those relationships. | Moderate | Public records; one conversation; no reference calls on network quality | The relationships are ceremonial rather than operational |
| This will probably work | My rough estimate is a 60% chance of achieving the one-year milestones, conditional on the crux checks passing; the estimate is a judgment, not a model output. | About 60%, judgment | Structured judgment; no base rate from comparable grants | Base rates from comparable fieldbuilding grants are much lower |

Non-trivial jump: This will probably work is not made transparent by adding a number. It becomes transparent when the writer says what work means, what the number is conditional on, and what kind of estimate it is. D01's examples show the same move: plausibly is used when the writer has a rough impression; a 70% interval is used when the writer has thought hard enough to give a range.

**Step 3**

- Score 2: the rewrite states a confidence expression, a kind-of-support phrase, and a crux for a load-bearing claim.

- Score 1: the rewrite states two of the three, or states all three for a non-load-bearing claim.
- Score 0: the rewrite substitutes an adjective, a citation without page or quote, or a number without a support phrase.
- Penalty: false precision. A confident-looking number with no model or base rate should be rewritten as a judgment with an explicit source of instability.
- Penalty: confidence theater. If the support is I talked to some people, the rewrite should say that, not dress it up as expert consensus.
- Penalty: citation laundering. A link without a page number or quote does not let the reader see the support.

Alternative defensible answer for some claims: do not quantify. D01 supports plausibly and seems likely when a claim is peripheral or when quantification would be unstable. A defensible rewrite of The field is neglected could be: I have not done a systematic landscape review; my impression from a shallow scan is that the field is thin, and I am not relying on this claim for the recommendation. That is more transparent than a fake 80% probability.

Alternative defensible answer for a high-stakes claim: provide a range and a model. If the cost-effectiveness claim is load-bearing for a multi-million-dollar grant, the rewrite should point to a BOTEC, state the key parameters, and give a range. D01's David Roodman example, raising a best estimate from 0.33% to 0.70% per decade and widening a 95% interval from 0.0-4.0% to 0.0-11.6%, shows what calibrated quantitative transparency looks like when the estimate is the output of a model.

**Step 4**

1. Internal memo variation: rewrite for a decision-maker who already shares the writer's worldview. Compress context; spend words on confidence, support, and cruxes.
2. External memo variation: rewrite for a skeptical reader who does not share the writer's premises. Add a where I am coming from paragraph and define terms that the internal reader could take for granted.
3. Short-timeline variation: rewrite for a decision that must be made in 48 hours. State the minimum viable confidence, the single crux, and the reversible next step. D01 does not require full secondary practices when the decision is time-sensitive and reversible.
4. Undocumented-conversation variation: rewrite when the support is a set of confidential conversations. State that the support cannot be fully summarized, say which claims rest on it, and give the reader a way to probe the crux without exposing sources.

5. Large-grant variation: rewrite for a $100 million scale grant. Add a research-process section, a data appendix, an archived source table, and a transcript or summary of expert calls where possible.

**Step 5**

- Confidence theater: adding probabilities without a model, base rate, or support phrase.
- Support laundering: describing a vague impression as expert input or field consensus without naming the expert or the basis.
- Buried crux: listing many risks while leaving the decision-relevant fact unstated.
- Overusing plausible: D01 notes the term can be interpreted differently by different readers. Use it for genuinely peripheral claims, not for load-bearing ones.
- Citation without locator: a link is not a quote or a page number.
- Data without code: a spreadsheet result without the underlying calculations is harder to audit.
- Transcript promise without delivery: if a conversation carries a major claim, either provide a summary or explain why not.
- Proportionality failure: applying a full GiveWell-style audit to a small reversible grant, or applying a rapid-grant memo to a large irreversible grant.
- Strategic-funder substitution: using transparency language to make the funder's vision look like evidence, rather than surfacing the grantee's local knowledge and the funder's uncertainty.

# Build a worked cost-effectiveness, BOTEC, expected-value and social-ROI comparison.

## Question

Build a worked cost-effectiveness comparison for two candidate grants, each sized at $500,000 so that cost does not do the work. Candidate Grant A is an effective-careers grant; Candidate Grant B is a bed-net-style global health grant.

The deliverable is a back-of-the-envelope calculation with assumptions, sensitivity, counterfactuality and caveats visible, plus an explicit statement of which single assumption dominates the comparison and where the ranking flips.

## Background

**Method anchors.** D01 supplies the writing discipline: open with a linked summary of takeaways, mark which considerations are most important, and state confidence and support for major claims. D02 supplies the BOTEC habit and the warning that an AI-generated back-of-the-envelope still needs review. L01 supplies the cost structure: work breakdown, cost pools that each have a single driver, and a KPI column tying spend to purpose.

**Numeric anchors.** P02 reports an Effective Careers RFP weighted mean social return on investment of roughly 5,500x, based on modelled estimates for about 85 percent of recommended grants, and explicitly calls these early and uncertain. P05 reports roughly 5x for the 2025 effective-giving RFP portfolio after quality and counterfactuality adjustments, and about 5-6x for the full portfolio. S01 reports a $212M bed-net grant expected to avert 67,000 deaths. P03 supplies the Coefficient Giving welfare convention.

**Design rule.** Every multiple below is stated with its numeraire, its counterfactual and its dominant parameter, and the sensitivity is shown before the recommendation. A high modelled ROI does not by itself establish that the marginal dollar was well spent.

## Sub-Question 1: Selecting the Numeraire and the Counterfactual

Before any arithmetic, two choices fix every number that follows: the numeraire (the unit in which benefit is measured) and the counterfactual (the state of the world against which the grant is compared). A multiple is meaningless without both.

*The corpus supplies three headline multiples that look comparable and are not. Treating them as one voice is the most common failure in this kind of exercise, so we fix the units first and compute second.*

**Step 1**

The three reported multiples answer three different questions, and the first non-trivial move is to refuse to compare them before aligning the object being valued.

Table 25: *Three multiples, three objects; none is a substitute for another.*

| Object valued | Numeraire | Reported or implied multiple | Source |
|---|---|---|---|
| Career transitions across the Effective Careers RFP portfolio | Modelled social ROI | ~5,500x | P02 |
| Dollars mobilised by the 2025 effective-giving portfolio | Quality- and counterfactuality-adjusted ROI | ~5x RFP; ~5-6x portfolio | P05 |
| Lives saved by the terminal health intervention | CG$ welfare units | ~1,040x (implied by this exercise) | S01 + P03 |

The 5,500x values a person's career path. The ~5x values dollars mobilised by an intermediary. The ~1,040x values lives saved by the terminal intervention. Ranking them against each other without aligning the object is a category error, and the corpus itself flags the hazard: P05 warns that marginal returns may fall as grantees scale, and P02 calls its own estimate early and uncertain.

```
# --- Anchors from the corpus, recomputed ---

# S01: AMF grant and expected deaths averted
amf_grant_usd = 212_000_000
amf_deaths_averted = 67_000
cost_per_death = amf_grant_usd / amf_deaths_averted

# P03: CG welfare unit. 1% income rise for one person for one year =
    CG$500.
# P03 also states the welfare gap (CG$6.1e15) equals ~1.86e9 lives.
cg_per_life = 6.1e15 / 1.86e9
cg_per_life_crosscheck = 381e12 / 115e6

b_welfare_per_usd = cg_per_life / cost_per_death

output_state = {}
```

```
output_state["cost_per_death"] = cost_per_death
output_state["cg_per_life"] = cg_per_life
output_state["cg_per_life_crosscheck"] = cg_per_life_crosscheck
output_state["b_welfare_per_usd"] = b_welfare_per_usd

print(f"cost per death averted (USD): {cost_per_death:,.0f}")
print(f"implied CG$ per life: {cg_per_life:,.0f} (cross-check
    {cg_per_life_crosscheck:,.0f})")
print(f"implied health-intervention multiple (CG$/USD):
    {b_welfare_per_usd:,.0f}x")
```

Two independent ratios inside P03 agree to within about one percent, which raises confidence that roughly 3.3M CG$ per life is the number the source actually implies rather than an artefact of my arithmetic. Note the provenance discipline: this value of a life is a derived cross-check, not a stated figure, and every CG$ result below inherits the contestedness of that welfare convention.

### Step 2

Expected value makes the weights explicit so that they can be argued with rather than assumed.

$$\mathrm{EV} = \sum_i p_i \, v_i, \qquad \mathrm{SROI} = \frac{\mathrm{EV}}{\mathrm{cost}}$$

SROI is just EV divided by cost; it is not a moral claim. Counterfactuality enters as a multiplicative factor, so a claimed transition count is inflated by definition until it is discounted to the transitions that would not have occurred anyway.

$$\mathrm{EV}_A = N_{\text{participants}} \cdot p_{\text{trans}} \cdot c \cdot V - \text{cost}$$

*The trap worth naming explicitly: two multiples can only be multiplied when the upstream multiple prices dollars mobilised and the downstream multiple prices welfare per dollar. P05 reports its ~5x after quality and counterfactuality adjustments, so value is already inside it. Multiplying by the terminal multiple again double counts.*

```
# --- Two different 'returns' and the trap of multiplying them ---

cost = 500_000

# Object 1 (intermediated): dollars MOBILISED per program dollar.
mobilization_multiple = 5.0   # P05: effective-giving RFP ~5x after
    quality+counterfactuality adj.

```

```
# Object 2 (terminal): welfare per dollar that reaches an effective
    health program.
terminal_multiple = input_state["b_welfare_per_usd"]

dollars_moved = cost * mobilization_multiple

# The NAIVE composite multiplies the two multiples. Correct only if
    the 5x is a pure
# dollar-mobilisation figure and does NOT already price terminal
    welfare.
naive_composite_multiple = mobilization_multiple * terminal_multiple

output_state = {}
output_state["dollars_moved"] = dollars_moved
output_state["naive_composite_multiple"] = naive_composite_multiple

print(f"dollars mobilized: ${dollars_moved:,.0f}")
print(f"naive composite multiple: {naive_composite_multiple:,.0f}x
    <-- double-count risk")
```

Adopt CG$ as the numeraire, state the counterfactual separately for each grant, and never multiply two already value-adjusted multiples. Both candidate grants below are therefore modelled from raw outputs to welfare, with the value parameters shown rather than buried.

## Sub-Question 2: Candidate Grant A: Effective Careers, $500,000

Grant A is $500,000 to a talent-advising organisation that moves participants into high-impact work. Model it bottom-up so that every input can be interrogated, and be explicit that the dominant input is the least anchored one.

### Step 1

Start from the simplest funnel that is still honest about counterfactuality.

```
# Candidate Grant A: $500k to an effective-careers / talent-advising
    nonprofit.
cost_A = 500_000

# Reach and conversion (illustrative; cross-checked against P07's HIP
    reference class)
participants = 500
p_transition = 0.06
gross_transitions = participants * p_transition

# Counterfactuality: share of transitions that would NOT have
    occurred anyway.
c_A = 0.4
```

```
net_transitions_A = gross_transitions * c_A

# P07 reference class: 51 career transitions from 210 IAP + 519 CIT
    participants
hip_rate = 51 / (210 + 519)

output_state = {}
output_state["cost_A"] = cost_A
output_state["participants"] = participants
output_state["p_transition"] = p_transition
output_state["gross_transitions"] = gross_transitions
output_state["c_A"] = c_A
output_state["net_transitions_A"] = net_transitions_A
output_state["hip_rate"] = hip_rate

print(f"gross transitions: {gross_transitions:.0f}")
print(f"net (counterfactual) transitions: {net_transitions_A:.0f}")
print(f"P07 HIP reference rate: {hip_rate:.3f}")
```

Twelve net transitions carry the whole of Grant A's impact. The reference-class cross-check from P07, where about seven percent of HIP participants transitioned in 2025, sits just above the six percent used here, so the conversion choice is not obviously self-serving. The counterfactuality factor of 0.4 is the more contestable number, and it enters linearly: halving it halves the multiple.

### Step 2

Now the load-bearing question: what is one counterfactual transition worth? Rather than pick a number, back-solve the number that P02's headline multiple implies.

$$V^* = \frac{\mathrm{SROI}_{\mathrm{target}} \times \mathrm{cost}}{N_{\mathrm{net}}} = \frac{5500 \times \$500{,}000}{12} \approx \$229\mathrm{M} \text{ per net transition}$$

```
cost_A = input_state["cost_A"]
net_transitions_A = input_state["net_transitions_A"]

target_sroi = 5500.0  # P02 weighted mean social ROI for the
    Effective Careers RFP
implied_value_per_transition = target_sroi * cost_A /
    net_transitions_A

V_A_central = 5_000_000  # illustrative central value per net
    transition, in CG$
EV_A_central = net_transitions_A * V_A_central
SROI_A_central = EV_A_central / cost_A
```

```

output_state = {}
output_state["implied_value_per_transition"] =
    implied_value_per_transition
output_state["V_A_central"] = V_A_central
output_state["EV_A_central"] = EV_A_central
output_state["SROI_A_central"] = SROI_A_central

print(f"implied value per net transition to reproduce 5,500x:
    ${implied_value_per_transition:,.0f}")
print(f"central EV_A: ${EV_A_central:,.0f} -> SROI
    {SROI_A_central:,.0f}x")
```

*This is the step that should stop the reader. To reproduce 5,500x from a dozen transitions, each one has to be worth roughly $229M in welfare units. That is not absurd inside a longtermist impact model, where a single well-placed person may shift the probability of a good outcome for the entire future, but it is an enormous claim and it is doing almost all of the work. My central value of $5M per transition is deliberately moderate and produces about 120x, not 5,500x. The discrepancy is the finding, not a defect in the arithmetic.*

### Step 3

Because the dominant parameter is a distribution rather than a point, compute the EV over explicit states with stated probabilities. This is hits-based giving made arithmetic.

```
cost_A = input_state["cost_A"]
n = input_state["net_transitions_A"]

# Hits-based: the portfolio is paid for by the tail. Distribution
    over VALUE PER
# net transition (CG$), with probabilities made explicit so they can
    be argued with.
value_states = [
    (0.50, 0.0),      # model was wrong / no realised impact
    (0.30, 5.0e5),    # modest
    (0.15, 5.0e6),    # central
    (0.045, 5.0e7),   # strong
    (0.005, 2.3e8),   # the value implied by a literal 5,500x reading
]
ev_V = sum(p * v for p, v in value_states)
EV_A = n * ev_V
SROI_A = EV_A / cost_A

output_state = {}
output_state["value_states"] = value_states
output_state["ev_V"] = ev_V
output_state["EV_A"] = EV_A
```

```
output_state["SROI_A"] = SROI_A

print(f"E[value per net transition]: ${ev_V:,.0f}")
print(f"EV_A: ${EV_A:,.0f} -> SROI {SROI_A:,.0f}x")
```

The tail term (0.5 percent at $230M) is what a literal reading of the 5,500x would require. Even then, the probability-weighted value is about $4.3M per transition and the SROI is roughly 103x. The model is dominated by the tail, and the tail is dominated by a single unobservable. A reader should ask what evidence could move that 0.5-percent weight, because nothing else in the calculation does much.

## Sub-Question 3: Candidate Grant B: Bed-Net-Style Global Health, $500,000

Grant B uses the cleanest anchor in the corpus. The exercise is to carry it to the same numeraire as Grant A without smuggling in a second valuation convention.

### Step 1

Derive the cost per death averted from S01 and scale it to $500,000.

```
cost_B = 500_000
cost_per_death = input_state["cost_per_death"]
deaths_averted_B = cost_B / cost_per_death

displacement_factors = [1.0, 0.75, 0.5]
deaths_by_displacement = {f: deaths_averted_B * f for f in
    displacement_factors}

output_state = {}
output_state["cost_B"] = cost_B
output_state["deaths_averted_B"] = deaths_averted_B
output_state["deaths_by_displacement"] = deaths_by_displacement

print(f"deaths averted (gross): {deaths_averted_B:.1f}")
print("by displacement factor:", {k: round(v, 1) for k, v in
    deaths_by_displacement.items()})
```

About 158 deaths averted for $500k. The displacement factors are not decoration: every figure here is gross until a counterfactual is named. P05 warns that returns fall as grantees scale, which is the same phenomenon as displacement seen from the funder's side rather than the beneficiary's.

### Step 2

Converting lives to a common unit is where the contestedness enters, and it is where the two grants finally become comparable.

$$\text{SROI}_B = \frac{D \times \text{CG\$}_{\text{life}}}{\text{cost}} = \frac{158 \times 3.28\text{M}}{0.5\text{M}} \approx 1036$$

```
cost_B = input_state["cost_B"]
deaths_averted_B = input_state["deaths_averted_B"]
cg_per_life = input_state["cg_per_life"]

welfare_B = deaths_averted_B * cg_per_life
SROI_B = welfare_B / cost_B

sensitivity = {cen: deaths_averted_B * cen / cost_B for cen in
    [1.0e6, 2.0e6, 3.28e6, 5.0e6, 1.0e7]}

output_state = {}
output_state["welfare_B"] = welfare_B
output_state["SROI_B"] = SROI_B
output_state["SROI_B_by_life_value"] = sensitivity

print(f"welfare_B: ${welfare_B:,.0f} CG$ -> SROI {SROI_B:,.0f}x")
print("SROI_B vs assumed CG$/life:", {f"{k:,.0f}": round(v) for k, v
    in sensitivity.items()})
```

The single number that drives $SROI_B$ is the CG$ value of a life, taken here from P03's own internal ratio rather than an imported value of a statistical life. At the central value Grant B returns roughly 1,040x, an order of magnitude above Grant A's central 103x. Both still sit on parameters that are not empirically pinned, so the result is provisional on the numeraire and on this conversion.

## Sub-Question 4: Stress-Testing and Deciding

Put both grants on one axis, find what flips the ranking, and then say whether the decision still turns on the unresolved parameter. If it does, the honest output is a conditional recommendation plus the evidence that would resolve it.

### Step 1

The point of the side-by-side table is not the ranking; it is that the ranking is entirely a function of the two dominant inputs, and each grant is fragile to a different one.

Table 26: *Same numeraire, same cost, different dominant parameter.*

| Grant | Cost (US$) | Central SROI (CG/$US$) | Dominant input | Break-even condition |
|---|---|---|---|---|
| Grant A (effective careers) | 500,000 | ~103x | Value per counterfactual career transition | Loses to B unless V rises above ~$43M |
| Grant B (global health) | 500,000 | ~1,040x | CG$ value of a statistical life | Loses to A only if CG$/life falls below ~$0.38M |

```python
cost_A = input_state["cost_A"]
cost_B = input_state["cost_B"]
SROI_A = input_state["SROI_A"]
SROI_B = input_state["SROI_B"]

comparison = {
    "Grant A (effective careers)": {
        "cost": cost_A,
        "central_sroi": SROI_A,
        "unit": "CG$ per US$",
        "dominant_input": "value per counterfactual career transition",
    },
    "Grant B (global health)": {
        "cost": cost_B,
        "central_sroi": SROI_B,
        "unit": "CG$ per US$",
        "dominant_input": "CG$ value of a statistical life",
    },
}

output_state = {}
output_state["comparison"] = comparison

for name, rec in comparison.items():
    print(f"{name}: {rec['central_sroi']:,.0f}x | dominant input: {rec['dominant_input']}")
```

At central values B dominates A by roughly a factor of ten. Treat that as a provisional result conditional on the numeraire and on the two dominant parameters, not as a settled empirical finding.

**Step 2**

Ask which assumption flips the ranking, not which grant wins. Reduce the comparison to a single shared axis: the value placed on one counterfactual unit of benefit.

$$V_{\text{flip}} = \frac{\text{SROI}_B \times \text{cost}_A}{N_{\text{net}}} \approx$$

\$43.2M per net transition

```python
cost_A = input_state["cost_A"]
cost_B = input_state["cost_B"]
net_tr = input_state["net_transitions_A"]
V_A = input_state["V_A_central"]
deaths_B = input_state["deaths_averted_B"]
cg_per_life = input_state["cg_per_life"]

SROI_A = net_tr * V_A / cost_A
SROI_B = deaths_B * cg_per_life / cost_B

# Single shared axis: value placed on one counterfactual unit of benefit.
V_flip = SROI_B * cost_A / net_tr            # A ties B at this value per net transition
cg_flip = SROI_A * cost_B / deaths_B         # B ties A at this CG$ value of a life

output_state = {}
output_state["SROI_A_recomputed"] = SROI_A
output_state["SROI_B_recomputed"] = SROI_B
output_state["V_flip"] = V_flip
output_state["cg_flip"] = cg_flip

print(f"A central SROI: {SROI_A:,.0f}x ; B central SROI: {SROI_B:,.0f}x")
print(f"A ties B when value per net transition = ${V_flip:,.0f}")
print(f"B ties A when CG$ per life = ${cg_flip:,.0f} (vs assumed {cg_per_life:,.0f})")
```

The break-even is stark. B loses to A only if the value placed on a life falls below roughly $0.38MCG$, well below any plausible anchor, or if the value per counterfactual career transition rises to about \$43M. The first is a low bar B clears easily; the second is exactly the number that has no external anchor.

*So the comparison is robust to the parameter we can constrain and fragile to the parameter we cannot. That asymmetry, not the size of the multiple, is the decision-relevant output.*

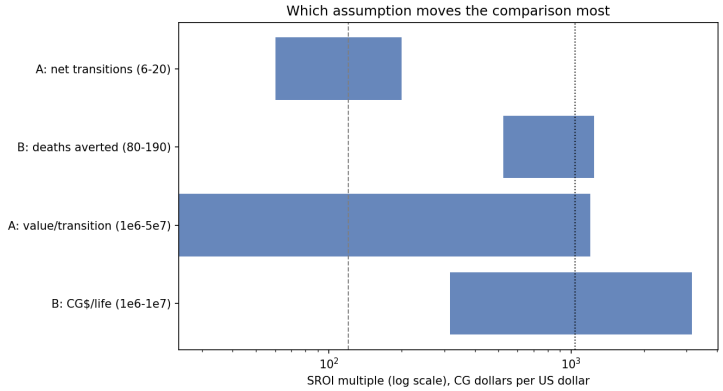


Figure 7: Graphical analysis for Step 2.

### Step 3

Expected value does not settle the decision. Counterfactuality, attribution and double counting all change the object being valued, and the marginal-dollar test asks a different question again.

- Counterfactuality: P02's 5,500x and P05's ~5x are both already value-adjusted; if the adjustment conventions differ, the numbers are not on the same footing.
- Attribution and credit: small early funders seeded organisations that later received large grants; counting the full terminal value to the later grant overstates it, an accounting cost the intermediary structure creates.
- Double counting: a career transition that routes someone into an effective-giving organisation can be counted at both ends of the same causal chain.
- Marginal versus inframarginal: the average multiple is not the marginal dollar's multiple, and P05 says marginal returns fall as grantees scale.
- Unit contestedness: P03's CG dollar is a contested quantification, and every CG dollar figure above inherits that contest.

```
n = input_state["net_transitions_A"]
cost_A = input_state["cost_A"]

# Stylised value-of-information UPPER BOUND on A's dominant parameter.
prior_states = [(0.7, 5.0e6), (0.3, 5.0e7)]
ev_prior = n * sum(p * v for p, v in prior_states)
ev_perfect = n * max(v for _, v in prior_states)   # perfect
    information picks the best state
evpi_upper_bound = ev_perfect - ev_prior

```

```
output_state = {}
output_state["EVPI_A_upper_bound"] = evpi_upper_bound

print(f"stylised EVPI upper bound on A's dominant parameter:
    ${evpi_upper_bound:,.0f} CG$")
print("Compare with a central EV_A of roughly $50M CG$ from the same
    parameter.")
```

This value-of-information figure is a stylised upper bound, not a plan: it assumes perfect resolution before commitment, which is exactly what the field cannot deliver. The practically relevant version is sequential. Front-loading small commitments preserves the option to scale later, which is the same logic P10 uses for early-stage funds and the same logic behind the one-off funding surge in the wider ecosystem.

- Fund B by default when the numeraire is CG dollar welfare, because it clears its break-even on the parameter we can constrain.
- Fund A only with an explicit, externally anchored value per counterfactual transition, not one back-solved from a headline multiple.
- Prefer the smaller, reversible commitment when the dominant parameter is unresolved and the option to re-fund is preserved.

*The stopping rule follows the writing account: model until the decision no longer turns on the unresolved parameter. Here it still does, so the honest output is a conditional recommendation plus the evidence that would resolve it, not a single number presented as the answer.*

# Analyze hits-based giving, power laws and portfolio construction.

## Question

What does it mean to run a grantmaking portfolio when the payoff distribution is heavy-tailed, and how should that change *(a)* how a single grant is evaluated, *(b)* how a portfolio is assembled and then read retrospectively, and *(c)* how grant size, stage structure and follow-on funding should be set?

Deliverable: a portfolio-construction memo that makes the hits-based logic explicit, separates **expected-value payoffs** from **observed hits**, names the selection risk built into retrospective success stories, and states what follows for grant sizes and follow-on funding.

## Background

Source P10 (EA Funds, Transformative AI Fund) states the doctrine directly: when funding early-stage work, expect **power-law-distributed payoffs**; most grants will yield nothing or little; what pays for the portfolio is the occasional home run. P10 also makes a second, separate claim: **most new organizations start with small checks**, a first grant of roughly $25k is often the difference between a pilot running or not, and many institutional funders do not serve that size for good reasons.

Source P11 (Manifund regranting) supplies the worked examples: Timaeus received roughly $143k in August 2023 and its later merger with Resolution sits upstream of Coefficient Giving's $160m grant, CG's largest AI safety grant; Manifund argues this one regrant alone would repay its $5.2m regranting budget to date. Neuronpedia, Joseph Bloom, WMDP, ChinaTalk and Apollo Research are also cited. *Apollo is flagged by Manifund itself as less exciting* because it had already taken roughly $2m from Open Philanthropy and SFF before the regrant.

Source P12 (Manifund, on the design of grant programs) decomposes any program into money, taste, dealflow, hustle and trust, and argues trust is the likely bottleneck for deploying a funding wave. P05 (Effective Giving & Careers) states the scale-side counter-principle: even mature grantees may not have hit a ceiling, marginal returns on additional dollars may be lower than in early stages, *but absolute gains from investing more in proven models remain large.* P15 (ACX Grants 2025) supplies the denominator discipline: 654 applications, 42 funded, with many good ideas unfunded for contingent reasons including the depth of evaluation required, conflicts of interest, and limited money. N01 is an unfunded event-organizing application and is explicitly not a success case.

The governing uncertainty for this step is the one the knowledge summary names: retrospective hit examples are **selected on the dependent variable**. They establish that a mechanism can operate; they do not establish base rates

or expected returns. Every step below must keep that distinction load-bearing rather than decorative.

## Sub-Question 1: What Hits-Based Giving Asserts, and What It Does Not

Before the claim can be applied, it has to be stated in a form that could be wrong. P10 asserts a shape for the outcome distribution and a consequence for how to fund. This sub-question produces the precise statement, the non-claims that people smuggle in, and the denominators that make the claim operational.

### Step 1

State the object being valued before stating the distribution. A grant portfolio has at least three distinct payoff objects and they do not share a distribution: dollars that return to the funder, social value created net of counterfactual, and option value in the form of information and follow-on rights. P10 and P11 are mostly talking about the second, but the arithmetic of the first is what the examples display.

$$\mathbb{E}[V] = \sum_{i=1}^{N} p_i \, v_i \qquad \text{vs.} \qquad \arg\max_i \; v_i$$

The non-trivial jump is here. Hits-based giving is *not* a decision rule of picking the largest possible outcome, and it is *not* a claim that probabilities can be ignored. It is a claim that the probability mass sits on rare large outcomes, so that **the expected value is dominated by the tail while the modal outcome is failure**. The two facts coexist; confusing them is the most common failure.

### Step 2

Four non-claims that must be separated from the doctrine, because each is a licence that the sources do not grant.

- It does not lower the bar. The bar is on expected value, which is probability-weighted. A grant with a 2% chance of a large outcome can clear the bar; a grant with a 2% chance of a small outcome cannot.
- It does not make weirdness evidence. Weirdness is a heuristic for neglectedness, a different property. A project can be neglected and still have low expected value; the two must be combined, not substituted.
- It does not license ignoring base rates. P15's 654 applications and 42 grants is the denominator fact that keeps hits-based reasoning honest. Without a denominator the doctrine degenerates into post hoc celebration.

- It is not a reason to expect failure in advance. The claim is about the shape of the outcome distribution, an empirical property that could be false for a given field and period. If realized outcome distributions are thin-tailed, the doctrine has no work to do.

**Step 3**

There are two denominators, and under a staff-time constraint they disagree. For a funder allocating money, the natural ratio is expected value per dollar deployed. For an organization whose binding constraint is judgment-hours, the natural ratio is expected value per hour of grantmaker attention. The prior phase established that CG describes itself as capacity-constrained, with staff time as the main obstacle to scaling impact, so the second ratio is frequently the operative one.

```
def ev_per_hour(p_hit, value_if_hit, hours):
    return p_hit * value_if_hit / hours

# ILLUSTRATIVE NUMBERS ONLY. These are invented for the exercise and
# do not represent any actual grant, applicant, or decision.
options = [
    {"name": "small_option_check", "p": 0.05, "v": 5_000_000,
    "hours": 4},
    {"name": "large_inframarginal", "p": 0.40, "v": 400_000_000,
    "hours": 400},
    {"name": "median_rfp_response", "p": 0.02, "v": 2_000_000,
    "hours": 20},
]

for row in options:
    ev = row["p"] * row["v"]
    output_state[row["name"] + "_ev"] = ev
    output_state[row["name"] + "_ev_per_dollar"] = ev / 1.0
    output_state[row["name"] + "_ev_per_hour"] = ev_per_hour(
        row["p"], row["v"], row["hours"]
    )
```

The interpretation is the point, not the arithmetic. A small option check can dominate on expected value per hour even while contributing almost nothing on expected value per dollar, and a large inframarginal commitment does the reverse. Under a binding staff constraint the portfolio is therefore **an optimization over two incommensurable ratios**, and a memo that reports only one of them is deciding covertly.

## Sub-Question 2: The Expected-Value Mathematics and Why Sample Means Mislead

Hits-based reasoning is only as good as the estimator underneath it. This sub-question makes the heavy-tail arithmetic explicit, shows numerically why an observed portfolio mean is close to uninformative, and specifies what would falsify the heavy-tail premise.

### Step 1

Parameterize the outcomes so the tail thickness is a dial rather than a mood. A lognormal payoff multiplier $r \geq 0$ with log-scale location and spread gives a tractable family in which one parameter controls how much of the expected value sits in the extreme upper tail.

$$\ln r \sim \mathcal{N}(\mu, \sigma^2), \qquad \mathbb{E}[r] = \exp\left(\mu + \tfrac{\sigma^2}{2}\right)$$

Two facts about this family carry the whole argument. First, the population mean is finite for every $\sigma$, yet the mean is governed by the exponential of $\sigma^2/2$: the sample mean converges at a rate that scales with $\exp(\sigma^2)$, so the required sample size grows explosively in tail thickness. Second, in the strict Pareto case with tail exponent $\alpha$, the variance is infinite for $\alpha \leq 2$ and the mean itself is infinite for $\alpha \leq 1$. **This is the mathematical reason why a report of the form 'we funded forty grants and they averaged 1.2x' carries almost no information**, and why retrospective mean-of-funded grants is not an estimate of expected value.

### Step 2

Simulate portfolios rather than single grants, because the portfolio is the level at which the claim has content. The script below reports three facts: the dispersion of the portfolio mean across replications, the share of realized portfolio value contributed by the single best grant, and the running sample mean for one replication to show how slowly it settles.

```
import numpy as np

rng = np.random.default_rng(7)

mu = float(input_state.get("mu", -3.0))
sigma = float(input_state.get("sigma", 4.0))
n_portfolio = int(input_state.get("n_portfolio", 500))
n_replications = int(input_state.get("n_replications", 5000))

draws = np.exp(rng.normal(mu, sigma, size=(n_replications,
    n_portfolio)))
means = draws.mean(axis=1)
```

```
top_share = draws.max(axis=1) / draws.sum(axis=1)

output_state["mean_of_portfolio_means"] = float(means.mean())
output_state["sd_of_portfolio_means"] = float(means.std())
output_state["median_portfolio_mean"] = float(np.median(means))
output_state["p95_portfolio_mean"] = float(np.percentile(means, 95))
output_state["median_share_from_top_grant"] =
    float(np.median(top_share))
output_state["median_grants_worth_nothing"] = float(
    np.median((draws < 1.0).mean(axis=1))
)
output_state["single_series_running_mean"] = (
    np.cumsum(draws[0]) / np.arange(1, n_portfolio + 1)
).tolist()
```

Read the output as three separate facts. The portfolio mean has a long right tail of its own: the mean of means sits far above the median, so a typical portfolio looks mediocre while a rare portfolio carries everything. The share of realized value from the single best grant is often large: at these parameters the median portfolio derives most of its value from one or two grants, which means the portfolio **is effectively a bet on whether at least one extreme draw landed inside it**. And the running sample mean does not visibly settle within one portfolio lifetime, which is the numerical form of the claim that you cannot learn the field's expected return from your own track record in a few years.

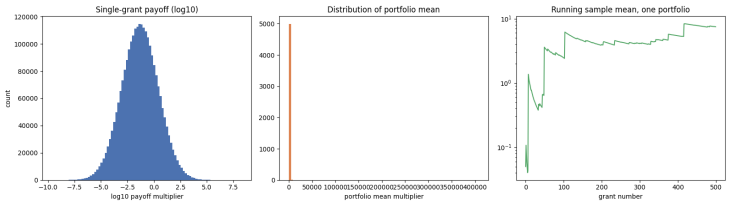


Figure 8: Graphical analysis for Step 2.

### Step 3

The premise is empirical and therefore falsifiable. The following observations would weaken or kill the hits-based case for a given field or program.

- A measured outcome distribution across a program's full funded set, including failures, whose tails are thin relative to the dispersion the doctrine assumes.
- Evidence that the largest outcomes are attributable to factors outside the funder's selection (timing, exogenous funding waves), which would compress the marginal effect of any individual grant.

- Evidence that the pool is dominated by a single correlated factor, so that the effective number of independent bets is near one and the tail cannot be reached by adding more grants.
- A binding constraint that is not money and not judgment-hours but organizational supply, in which case the relevant arithmetic is not portfolio payoff at all but the rate at which new organizations can be created.

Note what P10 actually does and does not supply. It supplies the doctrine; it does not supply an outcome distribution, a base rate, or an estimate of $\sigma$ for any of its portfolios. The falsifiers above are therefore open questions, not resolved ones, and a memo should label them as such.

## Sub-Question 3: Selection Risk in Retrospective Hit Lists

P11's argument is built from cases that were selected because they worked. This sub-question separates what each case can legitimately support from what it cannot, unpacks the credit-allocation and counterfactuality premises inside the headline multiplier, and states what evidence would repair the inference.

### Step 1

The structural problem is selection on the dependent variable. The cases were chosen because the outcome was large, so the sample cannot estimate the probability of a large outcome. A hit list is a set of existence proofs for a mechanism, not a sample from the population of grants. The table below reads each Manifund case through that distinction.

Table 27: Each row separates a mechanism existence proof from a base-rate estimate. The rightmost column is the discipline the memo must enforce.

| **Case** | **First check** | **Later outcome (per P11)** | **What it can support** | **What it cannot support** |
|---|---|---|---|---|
| Timaeus | $143k, Aug 2023 | Merged with Resolution; Resolution later received CG's $160m AI safety grant | A small first check can sit upstream of a very large later funding event | That the $143k caused the $160m outcome, or that the $160m belongs in the regrant's return |

| Case | First check | Later outcome (per P11) | What it can support | What it cannot support |
|---|---|---|---|---|
| Apollo Research | $335k, Aug-Dec 2023 | First empirical scheming evals; worked with OpenAI and Anthropic; grew to ~40 people; already had ~$2m from Open Phil and SFF | A regrantor can be early relative to an organization's lifetime | That the regrant was the marginal enabler; Manifund itself calls this the less exciting case |
| Neuronpedia and Joseph Bloom | $2.5k and $25k, Jul 2023 | Interpretability tooling used widely; Bloom later became model transparency lead at UK AISI | Very small checks can be decisive at the level of an individual | Base rates, since these are cases the program chose to feature |
| WMDP benchmark | $190k, Jan 2024 | Became the standard hazardous-knowledge benchmark in the unlearning subfield | A modest check can fund a widely used public artifact | That the artifact would not have been produced otherwise |
| ChinaTalk | $38k, Sep 2023-Aug 2024 | First outside funding; grew to five full-time staff; later funded by Emergent Ventures, Macroscopic and CG | First outside funding can precede growth and later institutional support | That this trajectory is typical of first checks rather than exceptional |

**Step 2**

The headline claim that a $143k stake grew **53x** to roughly $7.6m, repaying the program's entire $5.2m budget, contains several premises that are not the same claim. Austin Chen's list of philanthropic funding pitfalls names two of them: *unclear credit allocation* and *double-counting of impact*. Unpacking the multiplier into its premises is the non-trivial move.

- Premise 1, valuation not cash: the 53x is a mark derived from a much larger later grant, not a realized return. A mark can move in either direction and does not by itself repay anything.
- Premise 2, counterfactuality: the arithmetic implicitly attributes the later outcome to the earlier check. If a larger funder was going to make the grant anyway once the organization existed, the marginal effect of the small check is the acceleration, not the whole outcome.
- Premise 3, stage separation: if CG funds the later stage, the same dollars cannot be credited twice, once to the regrant as origination and once to the institutional grant as impact. Credit allocation is a design decision, not a fact.
- Premise 4, denominator: 'this grant alone repays the budget' is a statement about one case. It says nothing about the average expected value of the other grants in the same program, which is what a portfolio decision requires.
- Premise 5, endogeneity of selection: P11 also notes that Apollo had already raised roughly $2m, which is exactly the information a base-rate estimate would need and exactly the information a curated hit list tends to omit.

The argument for small checks survives the unpacking, but in a weaker and more useful form: the small check **buys an option and an information signal** at a stage where the later, larger decision is not yet available. That is a real value, and it is a different value from 'the small check produced the big outcome'.

**Step 3**

Case studies become usable when they are reconnected to a denominator and to a counterfactual. The repairs below are the ones a memo should commit to.

- Report the denominator alongside the hit: grants made, grants that produced nothing, and the total budget against which the hit is being claimed. P15's 654-to-42 ratio is the model for this, and N01 is the reminder that unfunded applications are part of the population.
- State the counterfactual for each hit in one sentence: what would have happened without the check. Where the honest answer is 'the organization would likely have existed anyway', say so and downgrade the attribution.

- Separate the two decisions being learned from. The first check decision and the follow-on decision are different skills with different evidence demands; a program can be good at one and bad at the other.
- Pre-register what a hit would look like before the round, so that the observed cases are not selected after the fact. This is the same move as P02's recorded forecast that fewer than 40 applications would arrive, which is valuable precisely because it was written down and was wrong.

## Sub-Question 4: Grant Size, Portfolio Shape, Follow-On Funding: The Memo

With the doctrine stated and the selection risk bounded, the remaining work is constructive: what portfolio shape follows from a heavy-tailed payoff distribution, how grant size and stage structure should be set, where value is actually captured, and what the memo should contain so that a reader can disagree with a specific step rather than with the whole frame.

### Step 1

The quantity that matters is not the number of grants but the number of ***independent*** shots at the tail. Independence is the hidden premise in every 'add more bets' argument, and it is usually false in an ecosystem where several bets depend on the same field-growth and funding-wave conditions.

$$P(\text{at least one hit} \mid n) = 1-(1-p)^n \quad \text{independent} \qquad \text{vs.} \qquad \rho\,[1-(1-p/\rho)^n] \quad \text{common shock}$$

```python
import numpy as np

p_hit = float(input_state.get("p_hit", 0.03))
rho = float(input_state.get("rho", 0.5))
n_bets = np.arange(1, 201)

p_independent = 1.0 - (1.0 - p_hit) ** n_bets

# Stylized common-shock model: with probability rho the shared condition holds,
# and conditional on it each bet hits independently with probability p_hit / rho.
# Marginally each bet then hits with probability p_hit.
p_conditional = min(p_hit / rho, 1.0) if rho > 0 else 1.0
p_correlated = rho * (1.0 - (1.0 - p_conditional) ** n_bets)

output_state["p_any_independent_at_100"] = float(p_independent[99])
output_state["p_any_correlated_at_100"] = float(p_correlated[99])
output_state["n_for_50pct_independent"] = int(
    np.argmax(p_independent >= 0.5) + 1
```

```
)
output_state["n_for_50pct_correlated"] = int(
    np.argmax(p_correlated >= 0.5) + 1
)
output_state["curve_independent"] = p_independent.tolist()
output_state["curve_correlated"] = p_correlated.tolist()
```

The model is stylized and is flagged as such. Its purpose is to make one point visible: correlation imposes a ceiling on how much diversification can buy. Beyond some number of bets, adding grants raises the probability of a hit only if the bets are genuinely independent; under a common shock the curve asymptotes at $\rho$, so the marginal grant is worth less than the average one. This is the arithmetic behind P12's claim that *trust is the likely bottleneck*: correlated bets also correlate on the interpersonal conditions that let money move at all.

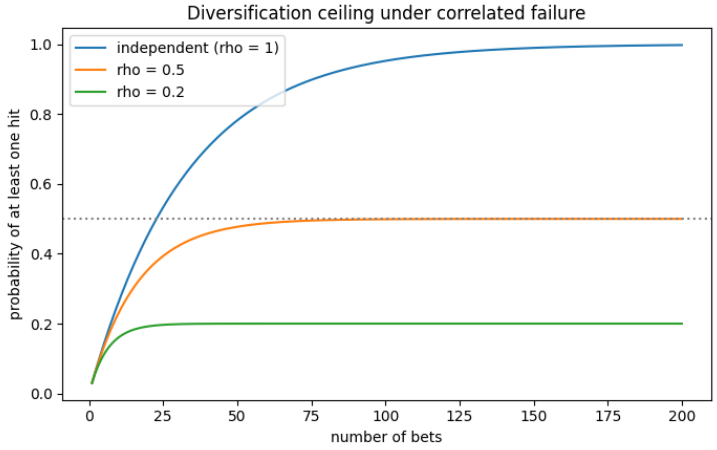


Figure 9: Graphical analysis for Step 1.

### Step 2

Grant size is not a single dial. Under a heavy-tailed distribution, small and large grants do different jobs, and the portfolio can rationally hold both at the same time. P05 supplies the scale-side argument that the marginal return on additional dollars to a proven grantee may be lower while the absolute gain remains large; P10 supplies the small-check side, that a first $25k is often decisive and that institutional funders do not serve that size. These are not in conflict; they describe two regions of the same portfolio.

- Small checks are options. Their value is the information they generate and the right to make a larger later decision, not the immediate impact of the funded activity.

- Large checks are inframarginal commitments. They are defensible when conviction is already high and the binding question is whether the organization can absorb the money well, not whether it should exist.
- Middle-sized grants carry the worst of both: too large to be costless experiments, too small to move the tail. They should be justified by an explicit reason rather than by default.
- The binding constraint may not be the funder's dollars. P10 argues the scarce thing is organizations, so a portfolio that ignores organizational supply will overstate what additional money can do.

**Step 3**

If small checks are options, the value of the portfolio is captured at exercise: the follow-on decision. This is the step the hit lists under-emphasize, because the follow-on grant is usually made by a different funder at a much larger size. Two consequences follow.

- Origination and scaling are different competencies. A regrantor's edge is dealflow and taste at the first-check stage; an institutional funder's edge is scale diligence and the ability to write a large check quickly. A memo should say which competency it is claiming and not average them into one.
- The follow-on decision is where credit is contested. P11's own Apollo note shows the mechanism working: an organization that had already raised roughly $2m is precisely the case where the marginal effect of a small check is smallest. Counterfactuality and stage are the same question.
- Capacity determines whether the option can be exercised in time. P10's argument for front-loading rests on the claim that the next 12-18 months disproportionately set the ecosystem's size; an option that cannot be exercised before the window closes is not worth much.

Caveat to state explicitly in the memo: the option framing assumes the signal from the small grant is informative about the later decision. If the information content is low, paying for it is a pure cost, and the small-check argument reduces to a claim about organizational supply rather than about learning.

**Step 4**

The deliverable is a memo. The table below gives the allocation units it must price, followed by the drill and the falsifiers that turn the memo from a description into a decision.

Table 28: Sizes are indicative and drawn from the source set: P13 and P14 for rapid grants, D04 for the typical project scale, P10 for the early-stage band, and P09 for the seeding tiers.

| **Allocation unit** | **Typical size** | **What it buys** | **Marginal-value logic** | **Falsifier** |
|---|---|---|---|---|
| Rapid / micro grants | up to about $10k | Concrete blockers: compute, travel for a specific deliverable, event costs | Very high expected value per judgment-hour; low absolute dollars | Funded work proceeds at the same rate without the grant |
| First checks / regranting | tens of thousands to low hundreds of thousands | Origination of new organizations and new projects; dealflow | Option plus information; value is realized at the follow-on decision | The follow-on funder was indifferent to the earlier signal |
| Project grants | roughly $200k to $2m per year | A defined deliverable with a theory of change | Moderate per-dollar; high per-hour when the track record is already legible | The deliverable would exist at similar quality without the grant |
| General support renewals | renewals, possibly larger than the initial grant | Continuity for proven organizations; room to raise ambition | Lower marginal return, large absolute gain, as P05 argues | The organization has hit an absorptive ceiling or drifted from its track record |
| New-organization seeding | $200k to $2m pre-seed, $2m to $20m seed | Organizations that should exist and do not | High labor per dollar at the start, larger later; the seed option | No credible founder can be found for the gap within the window |

| Allocation unit | Typical size | What it buys | Marginal-value logic | Falsifier |
|---|---|---|---|---|
| Large infra-marginal commit-ments | tens of millions and above | Scale for teams already above the bar by a wide margin | Dominated by the variance of one or two outstanding opportuni-ties | |

The allocation drill, stated as a task with two defensible answers. Take a portfolio of one hundred established grants and a marginal budget equal to one additional mid-sized grant. **Answer A, scale-biased:** allocate to the largest existing high-conviction grantee, on the ground that absolute gains from proven models remain large, that new organizations cannot be created on the timescale of the budget cycle, and that the marginal return may be lower but is still above the bar. **Answer B, tail-biased:** split the same budget into several first checks on new organizations, on the ground that the portfolio mean is dominated by the tail, that the mid-sized commitment is unlikely to be the marginal cause of anything, and that options must be created before they can be exercised. The drill is to say which answer the memo endorses **and what observation would move it to the other**.

- Memo section 1, the claim: the outcome distribution is heavy-tailed, stated with the parameters or qualitative bounds that make it falsifiable.
- Memo section 2, the estimator: why the sample mean of funded grants is not the quantity of interest, illustrated with the Monte Carlo above.
- Memo section 3, the evidence status: which claims are source-supported (P10's doctrine, P11's cases, P05's scale argument, P15's denominator) and which are inference (any specific expected return, any specific sigma, the allocation conclusion).
- Memo section 4, the selection risk: the dependent-variable problem, the credit-allocation premises, and the counterfactual for each cited hit.
- Memo section 5, the allocation: sizes, stage structure, the follow-on mechanism, and the two defensible answers above with the observation that would flip the recommendation.
- Memo section 6, the open questions: organizational supply versus dollars as the binding constraint; the independence assumption behind diversification; and whether the information from a first check is actually informative about the follow-on decision.

# Master grant-program design across money, taste, dealflow, hustle and trust.

## Question

Design a grant program, not only a grant. This step builds the program-level half of the evaluator's toolbox: a diagnostic instrument that locates the component a program is actually constrained by, and a worked comparison of two real programs across the five components that P12 names.

The deliverable has three parts: a five-component system model that replaces a checklist with a bottleneck structure; a diagnostic template with observable indicators and a decision rule, run once on a program from the corpus; and a head-to-head comparison of two programs with the design tradeoffs and failure modes that separate them.

## Background

Source-supported. P12 defines the five components: **money** to give out; **taste** in people, ideas and fields; **dealflow**, the funding opportunities that reach the funder; **hustle**, the legwork that coordinates money, taste and dealflow and then delivers the grant; and **trust**, a meta-component without which no grant is made. P12 also supplies the sharpest available comparison of programs, including ACX Grants, grantmaking.ai, Lightcone Commons and Coefficient Giving.

The corpus supplies two low-ticket, fast-decision programs to test the model against. P10 (EA Funds Transformative AI Fund) funds early-stage work with grants typically below $150k, rarely above $300k and almost never above about $1m, applying open applications, quick responses and transparency by default with privacy exceptions. P13, P14 and D03 (BlueDot Rapid Grants) fund concrete work already in motion, from $50 upward, up to $10k, with average replies inside a day and upfront payment by default.

Two disagreements are live and must not be averaged away. First, how much a funder should **shape** the ecosystem versus **respond** to it: P05 defends funding a shared service provider and ecosystem coordination, P06 warns against the strategic funder trap of slotting grantees in like subcontractors, P10 favours open calls that surface projects nobody thought to solicit, and P12 notes that much early-stage funding starts from a funder's own picture of what should exist. Second, whether direct grantmaking or funding intermediaries produces more impact, argued from opposite positions in P10 and P12.

Calibration note. The five-component vocabulary, the program descriptions and the figures are source-supported. The bottleneck formalisation, the principal-agent model, the diagnostic ratings and the tradeoff verdicts are *inferences* constructed for this exercise and should be labelled as such in an interview.

## Sub-Question 1: The five components as one system, not a checklist

P12 lists money, taste, dealflow, hustle and trust as the components any grant program needs to be good. The non-obvious move is to ask **what kind of object this list is**. Treat it as a production system with a binding constraint and a gate, not as five independent scores you would like to be high.

Explain how money, taste and dealflow interact to set the number and size of grants; explain what makes trust categorically different; and state the conditions under which adding dollars changes the number of grants, only the size of grants, or nothing at all.

### Step 1

The components, as defined by P12.

- Money: the pool to be given out. P12 traces the lineage from earning-to-give donations, through Dustin Moskovitz's money via Good Ventures, the FTX Future Fund boom, and now the anticipated Anthropic and OpenAI dollars.
- Taste: the ability to discern good from bad projects, in three forms that P12 separates: taste in people (founders, org leaders), taste in ideas (will this work, and would it be good if it did), and taste in fields (cause prioritisation across technical safety, policy and fieldbuilding). P12 flags the trap: everyone believes they have good taste, and there is no single true taste.
- Dealflow: the opportunities that reach the funder. Good dealflow means seeing projects other funders would not, seeing many more of them, being contacted first by good opportunities, or creating opportunities from scratch.
- Hustle: the legwork that coordinates money, taste and dealflow and delivers the grant. Not adversarial selling but polite persistence, and partly automatable, as Manifund does with digital grant agreements and self-serve payouts over Stripe Connect.
- Trust: a meta-component. If the other components do not trust each other, no grant is made. P12 expects trust, not capital, to be the bottleneck on deploying the coming funding wave.

Model the program as one gate bolted onto a non-substitutable core. Each grant needs capital, a taste decision, a deal to allocate to, and hustle hours to process; trust is a gate that can hold the whole thing at zero.

$$Q = \mathbf{1}\{\Theta \geq \theta_0\} \cdot \min\left(\frac{M}{a},\ T,\ D,\ H\right)$$

Here $Q$ is grants per year, $M$ committed dollars per year, $a$ the average ticket, $T$ appraisals per year, $D$ fundable deals surfaced per year, $H$ grant-operations throughput per year, and $\Theta$ the level of mutual trust, with a threshold $\theta_0$ below which nothing clears.

The consequence is the part worth memorising: money has **two margins**. When $M/a$ is the minimum and trust clears, extra dollars buy more grants. When it is not, extra dollars buy the same number of grants at a larger average ticket, because the binding component still fixes the count. That is exactly the arithmetic behind P01's larger grants and bigger average ticket, and the formal statement of P12's warning that money cannot always be converted directly into taste, dealflow, hustle or trust.

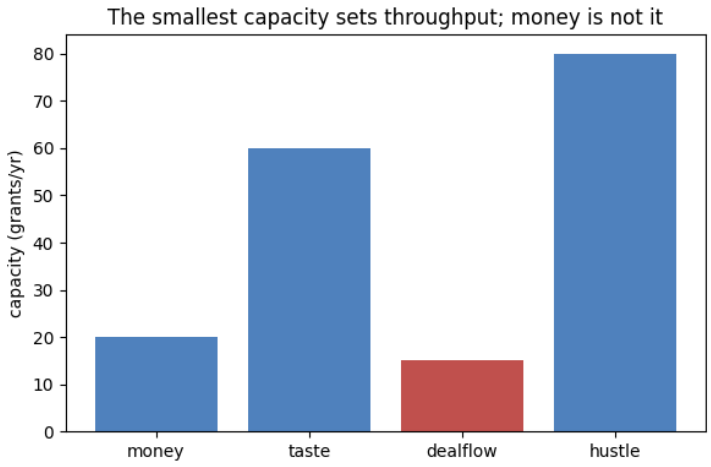


Figure 10: Graphical analysis for Step 1.

### Step 2

Trust is the gate, and P12 says it is where the coming funding wave will jam: the AI safety community started with high mutual trust because it was small, weird, demanding and socially interwoven, *and that probably will not scale*.

The failure it names is the cursed game of philanthropy: the principal who holds the money cannot verify that the agent who holds the taste is choosing well. Two structural fixes exist, and they trade against each other.

- Reduce the trust required. Lightcone Commons is described as placing final funding decisions in the hands of the money, so less trust in the evaluator is needed.
- Colocate money and taste. Manifund regranting and Leo Gao's micrograntsput the budget and the taste in the same person, so the agency gap

disappears for that unit.

$$EV_{\text{delegate}} = \rho\, V + (1-\rho)\, V_{\text{off}}, \qquad EV_{\text{colocate}} = \rho'\, V$$

$\rho$ is the probability the agent's chosen grant serves the principal's objective rather than the agent's own; $V$ is the value of the best grant the agent could find; $V_{\text{off}}$ is the value when the agent is misaligned. Colocation sets the misalignment term to zero but caps throughput at what one person can process, so $\rho'$ is bounded by the colocated person's capacity, not by their honesty.

Read the tradeoff directly: delegation beats colocation whenever $\rho$ is high enough, and the crossing point is where the extra reach of delegation stops compensating for the agency discount. In an interview, this is the argument that converts a soft word (trust) into a parameter you can reason about and ask about.

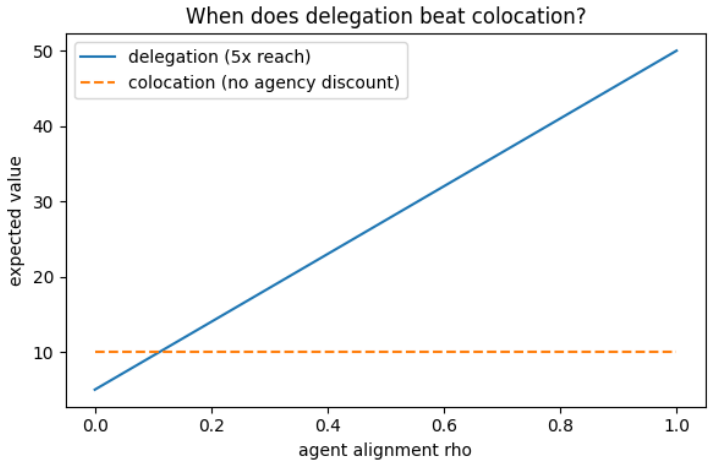


Figure 11: Graphical analysis for Step 2.

## Sub-Question 2: The diagnostic template: finding the binding component and the cheapest fix

Deliver a reusable instrument. It must turn the five-component model into questions a grantmaker can answer from artifacts already in hand, produce a rating, and return a diagnosis plus a proposed change.

The template must survive three hard cases: no single binding component, trust below its gate, and a program whose self-description hides the real constraint.

**Step 1**

Design principle: every diagnostic question must be answerable from material a funder already has, applications, budgets, public reports, response-time logs, or comparison with adjacent programs. A question the funder cannot answer from evidence is a question that will be answered by vibes.

Table 29: Diagnostic questions and their evidence base. The indicators are inferential; the components are P12's.

| **Component** | **Diagnostic question** | **Observable indicator** | **Failure mode if misread** |
|---|---|---|---|
| Money | Is capital, not judgment, the thing that stops the next grant? | Committed budget against the pipeline of fundable proposals; unused allocation | Assume money binds because it is the only component that is easily measured |
| Taste | Do you see more fundable projects than you can fund, and can you rank them? | Share of incoming proposals you would fund if unconstrained; disagreement among funders on the same file | Mistake confidence for calibration; a loud funder's taste substitutes for evidence |
| Dealflow | Do good opportunities find you, or do you have to create them? | Origin of the last ten grants: RFP, solicitation, inbound, or founder recruitment | Mistake brand reach for dealflow quality; hear only from organisations with comms teams |
| Hustle | How many staff-hours does one committed dollar cost? | Grants per staff-year; time from first contact to payment | Automate the easy step and leave the true bottleneck untouched |
| Trust | Do counterparties believe the money will arrive and be spent well? | Repeat grantees; referrals; co-funding; unforced disclosure | Buy trust with money and spend it on process; treat trust as binary when it decays |

Scoring and decision rule. Rate each capacity 0-3 against its own evidence, then apply the system model rather than a sum: the diagnosis is the minimum, unless trust is below its gate, in which case trust is the diagnosis regardless of the other ratings.

$$\text{diagnosis} = \begin{cases} \text{trust} & \Theta < \theta_0 \\ \arg\min_c \left(\text{capacity}_c\right) & \text{otherwise} \end{cases}$$

The change rule: propose the intervention that raises the binding capacity, then name the component it will push toward binding next. Every fix moves the constraint rather than removing the need for one, which is why the template outputs a **next binding component** as well as a diagnosis.

**Step 2**

Run the template on BlueDot Rapid Grants, using only published facts. D03 gives typical grants up to $10k, average reply within a day, nine in ten applicants hearing back within a week, payment upfront by default, eligibility for the BlueDot community and the wider AI safety and biosecurity community, and work already in motion with a specific bottleneck cost. P13 and P14 give the earlier $50 to $1,500 version, five-working-day targets, and the shift from compute-only to events, talent acceleration and community building.

Ratings, each against its evidence. Money 3 (tiny tickets against a large community surface; capital is not the limit). Hustle 3 (decisions in about a day, upfront payment, no formal report). Taste 2 (thin triage; the bar is concrete work in motion with a named bottleneck cost, not deep evaluation). Dealflow 2 (strong inside one community, thin outside it). Trust 2 (community eligibility substitutes for diligence, but only for those already inside).

Diagnosis. Trust clears the gate only because eligibility is restricted to a community BlueDot already trusts, which is the design's most interesting move: it converts a trust problem into an eligibility rule. That leaves **dealflow** as the binding capacity, because the program funds work that is already in motion, and that work must already exist. P14 is explicit: if you have not started, begin with what you have and come back when you hit a blocker.

One change. Add a **repeat-grantee fast lane**: pre-commit a small budget to any grantee who has completed a prior grant and filed a one-line update. This raises dealflow by converting prior grantees into a standing pipeline without paying a fresh trust cost, because trust was established by the first completed grant. Predicted effect: more grants per period at the same trust cost. Falsifier: if fast-lane grants are completed and reported at a lower rate than first-time grants, the trust transfer did not hold. Alternative defensible change: widen eligibility beyond the community, which trades trust for dealflow and should be paired with a slightly heavier verification step. Both answers are correct

depending on whether you believe BlueDot's brand or its community is the scarce asset.

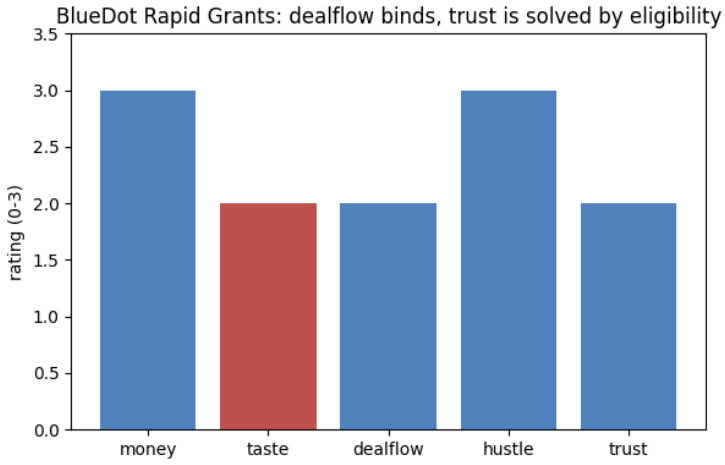


Figure 12: Graphical analysis for Step 2.

**Step 3**

Compressed script. A grant program is constrained by one of money, taste, dealflow, hustle or trust; trust is a gate and the others are non-substitutable. To diagnose one, ask which fails first when you try to make the next grant, rate each against artifacts you already have, take the minimum, and check the trust gate. Then propose the cheapest fix for the binding component and say what it will push toward binding next.

Limitations to state out loud. The ratings are ordinal and judgment-laden; two components can bind jointly; and the model treats trust as a gate when in reality it is continuous and decays. The template also cannot see a constraint outside the five, for example a legal or reputational cap on ticket size, or a mandate that excludes an entire cause area. If an interviewer presses, the right move is to name the missing component rather than defend the five as exhaustive.

## Sub-Question 3: Worked comparison: Coefficient Giving versus Manifund regranting

Compare two programs that P12 places almost at opposite corners of the same space. Coefficient Giving is the large institutional funder; Manifund regranting is a dispersed, part-time, colocated model. Describe each across money, taste, dealflow, hustle and trust, then state the tradeoffs and the failure mode each design invites.

The point is not to rank them. It is to see that each design buys one thing by giving up another, and to name what evidence would tell you which trade is right for a given funding problem.

**Step 1**

Coefficient Giving, from P06, P01, P11 and P12. Money is mostly Good Ventures with external donors more than doubling to over $200m. Taste is an army of specialised grantmakers spanning sub-funds. Dealflow is a mix of RFPs and active grantmaking, and because the funder is large, essentially all major organisations seek it out. Hustle is a second army of operations and logistics staff. Trust was originally inherited through GiveWell and now stands as an institution in its own right.

Manifund regranting, from P11 and P12. Money is modest by comparison, a $2m pool in 2025 moving toward a $5m target, with donors asked to cover five percent of operating cost. Taste and dealflow are dispersed across a pool of about 13-15 part-time regrantors with track records, each with a pot of $100k to $200k and top-ups for good grants. Hustle is largely automated through digital grant agreements and self-serve payouts. Trust rests on individual regrantor track records plus Manifund's transparency commitment.

Table 30: P12 supplies both rows directly; the 'what differs' column is inference.

| **Component** | **Coefficient Giving** | **Manifund regranting** | **What actually differs** |
|---|---|---|---|
| Money | Over $1b institution-wide in 2025, mostly Good Ventures | About $2m in 2025, targeting $5m | Scale versus agility of the capital pool |
| Taste | Distributed across specialised grantmakers | Colocated with the money in part-time regrantors | Division of labour versus colocation |
| Dealflow | RFPs plus active grantmaking; inbound at scale | Niche and personal; finds projects before centralised funders | Breadth of reach versus depth of contact |
| Hustle | Staffed operations and logistics | Automated agreements and payouts | Headcount versus tooling |
| Trust | Institutional, inherited via GiveWell | Personal track records plus disclosure | Banked reputation versus earned per regrantor |

The table hides the interesting part: the components substitute across rows. CG compensates for thin trust in any one grantmaker with institutional trust; Manifund compensates for thin money with colocated taste and automated hustle. This is why neither list of five can be scored independently.

**Step 2**

Two metrics separate the designs. First, throughput: dollars committed per grantmaker-hour. CG minimises staff per dollar on large tickets; P06 reports it moves more money and more grants per staff member than comparable philanthropies, and P01's larger average tickets are explicitly an attack on the staff constraint. Manifund gets a second form of throughput: it scales by adding part-time regrantors rather than by hiring and managing, which P11 argues is far lower friction.

Second, option value: the value of a portfolio of small, early bets under power-law payoffs. P10 states it plainly, most of these grants yield no or modest returns, and what pays for the portfolio is the occasional home run. That changes what good means for each program: CG is judged on its large grants, Manifund on whether any small grant becomes one.

$$\mathbb{E}[\text{portfolio}] = \sum_{i=1}^{N} p\, V_i$$

Take N early bets, each with probability $p$ of becoming a hit worth $V_i$, with the values dominated by the tail. A colocated regrantor is a cheap option on that tail precisely because the decision costs almost nothing to make. That is the structural reason Manifund can exist at all, and the reason CG cannot simply copy it: a $100k central review process would destroy the option.

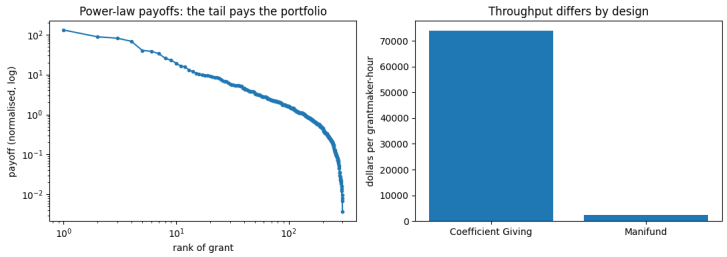


Figure 13: Graphical analysis for Step 2.

**Step 3**

The tradeoffs, stated as trades rather than preferences.

- Scale versus contact. CG's reach means all large organisations approach it, which is excellent dealflow for large tickets and poor dealflow for things that do not yet exist. Manifund's contact means it hears about projects before they are legible to a large funder, which is excellent for small tickets and cannot fund anything large.
- Colocation versus division of labour. Putting money and taste in one regrantor removes the agency discount and is cheap to scale by adding people; a large institution can specialise taste across many domains in a way no lone regrantor can.
- Institutional trust versus personal trust. CG can spend in a domain where it has never personally earned trust, because the institution carries it; Manifund must earn trust per regrantor, which is why it screens regrantors hard, and why its trust does not automatically transfer to a new hire.
- Throughput versus diversity of bets. A staff-hour constraint pushes toward larger tickets; P12's observation that everyone believes they have good taste pushes the other way, toward more independent deciders.

Failure modes, mapped to the design.

- Coefficient Giving. Bureaucratic drift and risk aversion as it scales, which P06 and S08 name explicitly; the strategic funder trap of thinking it has the answers and slotting grantees in like subcontractors; and slow response times when many funders compete, which S10 calls the game of chicken between funders leading to glacial response times.
- Manifund regranting. Power-law payoffs mean most grants return nothing, so the program lives or dies on a few hits; selecting good regrantors is itself a taste problem with no external check; colocation cannot fund anything above the small-ticket scale; and credit and counterfactuality across the small-to-large handoff are genuinely contested, as the Timaeus-to-Resolution sequence in P11 shows.

What each design protects against. CG's institution insures against its own grantmakers being wrong about a domain; Manifund's dispersion insures against the institution being wrong about a field. The falsifier for thinking you have chosen correctly is behavioural: if you never fund outside the people you already know, you bought the other model's weakness.

## Sub-Question 4: Live disagreements and the evidence that would settle them

The five-component model contains contested design choices, and the corpus does not resolve them. State each disagreement, the strongest case on both sides, and the concrete observation that would move a decision.

Do not merge the voices. P10 and P12 argue for intermediaries and colocation from inside organisations that practise them; P05 and P06 argue for shaping

and for marginalism from inside CG; P16 and P17 argue about nonfinancial grants from a microgrant program.

**Step 1**

The disagreements as a decision table. Each row is a real fork where two source-supported positions point different ways, and each row ends with the observation that would move you rather than a verdict.

Table 31: Positions are source-supported; the framing of each fork as a two-sided choice is inference.

| **Design fork** | **Case for side A** | **Case for side B** | **Observation that would move you** |
|---|---|---|---|
| Shape or respond to the ecosystem | P05 funds a shared service provider such as Good Impressions, because organisations will not buy services they have not tried; a funder can create a gap that should exist | P10 prefers open applications because a funder's view of what should exist misses projects nobody thought to solicit; P12 warns that much early-stage funding starts from the funder's own picture | Track where the good grants originate: solicited by the funder versus surfaced by an open call |
| Fund directly or fund intermediaries | P12 and P10 argue funds beat most solo donors on dealflow, expertise and operations, and have real edges | P12's trust argument: colocated regrantors remove the agency discount and can find things before centralised funders | Compare dollars-per-hit and counterfactuality-adjusted hit rate for the same field |

| Design fork | Case for side A | Case for side B | Observation that would move you |
|---|---|---|---|
| Transparency by default or privacy by exception | P10 publishes quarterly reports by cause area and grant size because disclosure builds trust and dialogue with donors and grantees | P10 also withholds grants whose work must stay private, listing them only in aggregates | Whether disclosure has ever cost a grantee a hire, a partner or a funder |
| Nonfinancial grants or money only | P17 notes several grantees said recognition helped them more than cash | P17 also notes that if a seal of approval is free it trends to being given to everyone, destroying its signal value | Whether a nonfinancial grant changes an applicant's outcome relative to a cash grant of the same cost |
| How much evidence a small grant needs | BlueDot rapid grants decide in about a day on work already in motion (D03, P13) | CG's AI governance RFP prefers first grants of $200k to $2M per year over one to two years (D04) | Cost per unit of decision-relevant information for small versus large tickets |

Two entries deserve expansion because they are the most likely to be probed.

Counterfactuality and credit. P11 reports that a $143k regrant to what became Timaeus grew into an eventual $160m CG grant, and asks whether that makes all of the regranting budget worth it. S10 lists unclear credit allocation and double-counting of impact among philanthropy's pitfalls. The honest position is that the two funders should not both count the full downstream impact; the question is how to apportion it, and the corpus does not answer it. Saying this out loud is a calibration signal, not a weakness.

Direct versus fund giving. P10 argues funds win on dealflow, expertise and operations for most individual donors, while granting that a well-informed solo donor may beat a fund. P12's five-component frame is the real reply: a fund is a way to buy taste, dealflow and hustle you do not have, at the price of trusting someone else with the money. That reframes the question from which is better to **which component are you short of**. If you are short of money, a fund is useless; if you are short of dealflow, it is the whole answer.

**Step 2**

Make the small-versus-large evidence dispute quantitative. A large grant justifies hours of diligence because the diligence cost is small relative to the stake; a small grant cannot, so it must either pre-accept a higher error rate or move the evaluation cost outside the grant, as an eligibility rule or a community does.

$$h \cdot c < p_{\text{error}} \cdot V, \qquad h^* = \frac{p_{\text{error}}\, V}{c}$$

$h^*$ is the diligence-hour budget, $c$ the fully loaded cost of an evaluator hour, $p_{\text{error}}$ the chance diligence changes the call, and $V$ the stake. Read it as a schedule: a \$10k grant with a \$200 per hour evaluator buys single-digit hours before diligence costs more than the error it prevents. The regime switch is the whole design implication: below roughly \$10k the honest cost of an extra hour exceeds the expected error it prevents, so a program must substitute a **pre-committed eligibility rule** for per-grant diligence. BlueDot's community rule and Manifund's track-record regrantors are exactly that substitution, viewed from the trust component.

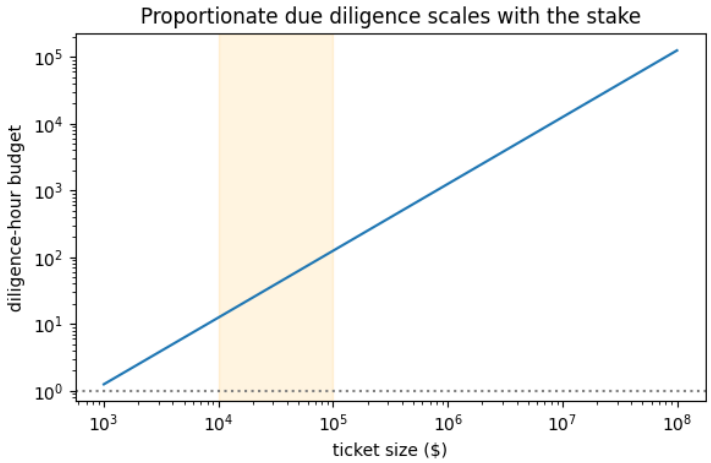


Figure 14: Graphical analysis for Step 2.

# Set proportionate due diligence, evidence thresholds and investigation tiers.

## Question

Active step: set proportionate due diligence, evidence thresholds and investigation tiers. Deliverable: a tiered due-diligence checklist keyed to grant size, risk and uncertainty, with red flags, decisive-evidence tests and a stop rule.

## Background

**Source-supported anchors: the evaluator brief calls for proportionate due diligence and for judging people and organizations. D04 lists assessment criteria as theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale; it says the expression-of-interest stage primarily assesses the first three; it typically prefers first-time grants of $200k-$2m per year over one to two years, with renewal after evidence of success; and it says projects below $200k per year should ideally be funded by other funders, naming SFF, LTFF, Founders Pledge, AI Risk Mitigation Fund and Longview. P04 describes asking what evidence would change one's mind, testing assumptions, checking track record with colleagues, assessing theory of change and downsides, and making timely decisions. L01 explains justifying budgets through cost drivers, cost categories tied to work packages, KPI contribution, prior spending, reference classes and mechanisms. D01 requires a linked summary of key takeaways, the most important considerations, and confidence plus support for major claims.**

*Inferences and limits: CG's internal risk tolerance and red lines are not fully public; D04 is program-specific; and the tier boundaries below are proposed craft, not a CG-published scoring rubric. Treat the thresholds as a rehearsal instrument and ask the interviewer for the current actual bar.*

## Sub-Question 1: The Proportionality Principle: What Varies and What Does Not

Due diligence is an information-acquisition problem under a budget. The target is not maximum evidence; it is enough evidence to make the decision as well as the stakes and reversibility warrant. The role brief supplies the norm: proportionate due diligence.

*Three axes vary independently: ticket size, downside or risk severity, and uncertainty. A non-trivial jump is that ticket size is a weak proxy for diligence. A small dual-use biosecurity grant can deserve more scrutiny than a large renewal to a proven global-health charity. Source anchors: role brief, D04, P04.*

**Step 1**

Define the decision and the cost of error. Ask: what is the maximum plausible downside if this grant is wrong? If it is small and reversible, light diligence is appropriate. If it is irreversible, catastrophic, or reputational, heavy diligence is appropriate. This is the risk axis.

Define uncertainty. Is the key claim already established by independent evidence, or is it a founder's projection? High uncertainty with tolerable downside calls for staged funding or a decisive-evidence test, not necessarily a longer memo.

Define ticket size. Ticket size sets approval level, expected value of information, and the number of people who need to sign off, but it does not by itself determine the required depth for every criterion.

**Step 2**

D04 evaluates applications holistically. The expression-of-interest stage focuses primarily on theory of change, track record, and strategic judgment. Full review examines these more thoroughly alongside project risks, cost-effectiveness, and scale. This describes changing emphasis, not three automatic rejection gates.

- Initial emphasis at EOI: theory of change, track record, and strategic judgment. Identify which uncertainties matter for the decision before choosing further investigation; D04 also accepts evidence of promise for future success, not only a matching past track record.
- Full-investigation criteria: project risks, cost-effectiveness, and scale, alongside deeper examination of the first three considerations. Evaluate the combined case rather than treating an earlier screening decision as conclusive.
- *Proposed practice: choose the next check by its expected decision value. An early rough calculation can sometimes reveal a decisive issue; strong track record does not replace project-risk analysis. This investigation sequence is a recommendation, not a published D04 requirement.*

**Step 3**

Make-or-refer boundary. D04 says first-time grants are typically $200k-$2m per year over one to two years, with renewal after evidence; below $200k per year should ideally be funded by other funders. P10 says the Transformative AI Fund typically funds below $150k, rarely above $300k, and no grants above roughly $1m for the foreseeable future. P13 and D03 describe BlueDot Rapid Grants up to $10k. These are not one universal CG threshold, but they establish that proportionality includes a routing rule: the correct action for a below-threshold grant may be referral, not deeper investigation.

## Sub-Question 2: Tiered Due-Diligence Checklist Keyed to Size, Risk, and Uncertainty

Build a tier matrix and checklist. Tiers are illustrative, anchored to source thresholds where available. The key is that each tier specifies a different evidence threshold, decision owner, and stop condition, not merely a longer form.

**Step 1**

Table 32: *Tier boundaries are illustrative and anchored to source thresholds where available. The tier is a routing decision, not a measure of applicant worth.*

| **Tier** | **Typical scale and examples** | **Primary failure mode** | **Evidence threshold** | **Core artifacts** | **Timeline** | **Decision and stop rule** |
|---|---|---|---|---|---|---|
| **Tier 0 Rapid** | Up to $10k; BlueDot Rapid Grants; Manifund microgrants; conference or compute support. | Wasted small dollar; nonexistent project; no bottleneck; fraud or sanctions. | Concrete work already in motion; specific cost is the bottleneck; reason the work matters for AI safety or biosecurity; eligibility; no sanctions or integrity red flag. | Short application; link to work; budget line; one reference only if needed; proof of prior spend if reimbursement. | Same day to five working days; one program officer can decide. | If yes, pay upfront where possible. If no artifact or cost is not the bottleneck, decline. If ambiguous, ask one clarification question; do not open a full investigation. Source anchors: P13, P14, D03. |

| Tier | Typical scale and examples | Primary failure mode | Evidence threshold | Core artifacts | Timeline | Decision and stop rule |
|---|---|---|---|---|---|---|
| **Tier 1 Small** | About $10k-$150k; early-stage projects; Rapid Grants scale-up; Manifund regranting; Transformative AI Fund typical zone. | Weak project; founder not ready; counterfactuality failure; small grant that should be a pilot. | Theory of change in one page; track record or strong advisor; budget with cost drivers; counterfactual reasoning; light external reference. | Application; budget; one or two references; written clarification; no full legal due diligence unless foreign or for-profit. | One to four weeks. | Fund as a pilot or milestone-based grant. Stop if the decisive evidence cannot be obtained within the pre-committed time. Source anchors: P10, P11, P12, N01-N06. |

| Tier | Typical scale and examples | Primary failure mode | Evidence threshold | Core artifacts | Timeline | Decision and stop rule |
|---|---|---|---|---|---|---|
| **Tier 2 Medium** | About $200k-$2m per year over one to two years; D04's typical first-time grant range. | Plausible but wrong theory of change; execution risk; hidden financial or governance problem; cost-effectiveness overstated. | D04 criteria: theory of change, track record, strategic judgment, project risks, cost-effectiveness, scale. Strategic judgment test: can the applicant name what would change their mind, prioritize risks, and reason under uncertainty? | Full proposal; references; reconstructed BOTEC; budget and WBS; financial controls evidence; grant write-up with linked summary. | Two to six weeks. | Fund first tranche; renew or top up only after evidence of success. Stop if two or more fatal red flags appear or if a load-bearing claim remains unsupported after the investigation budget. Source anchors: D04, P04, L01, D01. |

| Tier | Typical scale and examples | Primary failure mode | Evidence threshold | Core artifacts | Timeline | Decision and stop rule |
|---|---|---|---|---|---|---|
| **Tier 3 Large** | Above $2m per year or above $10m total; Resolution-scale, institute-scale, multi-year renewals. | Large irreversible commitment to the wrong team or strategy; governance failure; legal or compliance failure; reputational or safety downside; inability to absorb. | Independent outcome evidence; leadership and board references; external expert red-team; scenario and downside analysis; financial audit or equivalent; legal due diligence; co-funder coordination; milestone design. | Full investigation memo; independent references; risk register; legal review; staged grant agreement with gates; sign-off beyond the primary investigator. | One to three months, with fast-track only when clear evidence exists. | High conviction required; bet big only from a position of strength. Stop if any fatal legal, safety, governance, or integrity flag appears; if downside is catastrophic and unmitigated; or if the organization cannot absorb the grant. Use staged funding if uncertainty remains. Source anchors: D04, P01, P04, P06, D01. |

| Tier | Typical scale and examples | Primary failure mode | Evidence threshold | Core artifacts | Timeline | Decision and stop rule |
|---|---|---|---|---|---|---|
| **Tier 4 Proactive seed** | Pre-seed $200k-$2m; seed $2m-$20m; scale $20m-$200m+; Project Tailwind. | No founder; wrong founder; organization that should not exist; strategy that cannot attract others; CG overshaping. | Founder quality and agency; clarity about the unowned problem; theory of change; ability to recruit; counterfactuality; strategic importance; willingness to be wrong. | Founder diligence; references from people who have worked with them; cofounder plan; staged milestones; option to incubate. | Rolling; can be fast if the founder is known and the problem is clear. | Fund the person and the problem, not a fully specified plan. Stop if the founder is not exceptional and the problem is not clearly important; do not seed merely to fill a portfolio. Source anchors: P01, S09, P06. |

**Step 2**

Checklist by tier. The checklist is deliberately redundant across tiers so that a tired investigator can run it quickly. The point is not to complete every item; the point is to know which item is decision-relevant at this tier.

- Tier 0 checklist: confirm work in motion; confirm cost is the bottleneck; confirm eligibility; screen sanctions and integrity; record decision rationale in one paragraph; pay or reimburse.

- Tier 1 checklist: one-page theory of change; budget with cost drivers; counterfactual; one reference; simple risk list; pilot milestone.
- Tier 2 checklist: D04 six criteria; BOTEC; WBS or CBS; references; financial controls; risk register; grant write-up.
- Tier 3 checklist: independent outcome evidence; leadership references; external red-team; legal and compliance; co-funding; staged agreement; decision memo.
- Tier 4 checklist: founder references; unowned problem; strategic fit; staging; incubator option; exit criteria.

**Step 3**

Table 33: The overlay converts size, risk, and uncertainty into a tier adjustment. The most common correction is to bump a small but dangerous grant upward.

| **Condition** | **Tier adjustment** | **Rationale** |
|---|---|---|
| High downside severity, such as dual-use, biosecurity, model-weight security, or jailbreak tooling. | Bump one or two tiers regardless of size. | Risk axis dominates ticket size. Small grants can create large harms or reputational exposure. |
| High uncertainty with low downside. | Keep tier but use staged funding or a decisive-evidence test. | Buy information with a smaller first check rather than a larger memo. |
| Strong track record with the same team and same work. | Reduce project-level diligence, not financial or legal diligence. | Track record substitutes for some project risk, not for governance. |
| Outbound or proactive opportunity. | Tier 4. | The opportunity does not exist yet. Evidence is about people and strategy, not current artifacts. |
| Below CG comparative advantage and outside scope. | Refer, do not investigate deeply. | D04 says below $200k per year should ideally be funded by other funders. |

## Sub-Question 3: Red Flags, Decisive-Evidence Tests, and the Stop Rule

The checklist becomes operational only when it specifies what would stop an investigation and what evidence would change the decision. Use D01 reasoning transparency and P04's evidence-seeking habit: ask early what piece of evidence would most change your mind, then seek it time-efficiently.

### Step 1

Table 34: Illustrative investigation heuristics, not D04 rejection rules. Match verification to the material uncertainty, and distinguish evidence gaps from credible severe-risk or legal barriers. D04 calls for holistic review and lawful funding.

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Theory of change is a list of activities, not a causal chain, and has no falsifier. | You cannot tell whether the grant works, why it works, or what would update you. | Proposed heuristic: clarify the causal links, key assumptions and evidence that would weaken them. Missing articulation is an evidence gap, not automatic rejection by tier. Verify disputed links in proportion to their importance for the decision, then assess the combined case. | D04, D01, L02. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Track record mismatch: the organization has never done this work and has no relevant prior outcome. | Execution risk is not evidence-based. | Proposed heuristic: assess task-relevant ability and evidence of promise. D04 allows some applicants without matching track records. Higher execution risk can justify additional checks, relevant advisors or staged commitments; neither mandatory advisor-plus-staging conditions nor automatic Tier 3 rejection are D04 rules. | D04, P04. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Founder or key person is spread thin and there is no second-in-command. | Principal-agent and capacity risk; P04 explicitly names this as a common biggest risk. | Proposed heuristic: verify realistic workload, competing commitments and continuity arrangements. Request a proportionate capacity plan where delivery depends on one person. A missing deputy is not an automatic disqualification; unresolved capacity concerns should inform the overall execution assessment. | P04. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Budget lacks cost drivers; it lists resources but does not explain unit costs. | You cannot judge cost-effectiveness or value for money. | Proposed heuristic: request the quantities, unit costs or other assumptions that materially determine budget reasonableness. Scale the detail to the decision and investigation stage. Missing cost drivers reduce confidence and may require clarification; D04 does not make them an automatic Tier 2 rejection. | L01, D02. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Strategic judgment failure: the applicant cannot name what would change their mind, cannot prioritize risks, or cannot reason about counterfactuals. | The applicant is likely to fail under uncertainty and is a poor fit with D04 criteria. | Proposed heuristic: test strategic judgment through concrete choices, tradeoffs and uncertainty. Distinguish an unclear interview answer from repeated evidence of poor decisions; seek corroboration when consequential. Persistent, material weaknesses can weigh heavily in holistic review, without becoming an automatic tier-based veto. | D04, P04, L02. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Project risks are not identified or are dismissed, and the downside is catastrophic or irreversible. | Risk severity dominates expected value. | Proposed heuristic: pause support for the risk-bearing activity when a credible catastrophic or irreversible downside remains unresolved; seek specialist assessment of likelihood, scale, mitigations and alternatives. D04 does not require zero risk, so severity alone is not an automatic rejection rule. | D04, D01. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Cost-effectiveness claim has no denominator, no counterfactual, or double-counts impact. | The social return estimate is not decision-relevant. | Proposed heuristic: clarify the outcome unit and counterfactual, and correct double-counting before relying on the estimate. Test whether corrections could change the comparative case. An unusable estimate is not by itself a tier-based rejection; assess the remaining evidence and uncertainty explicitly. | P02, P05, P11, S10. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Financial controls are weak; there is no audit, no runway, or related-party transactions are undisclosed. | Governance failure can destroy the grant and trust. | Proposed heuristic: distinguish absent audits or limited runway from substantiated control failures or undisclosed conflicts. Verify records, material transactions and safeguards in proportion to the exposure. Escalate credible misuse or compliance concerns; neither a missing audit nor a tier label alone establishes disqualification. | P01, D04. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Legal or compliance problem: sanctions, foreign-funding restrictions, or charitable-status issues. | CG will only recommend funding when it is able to do so consistent with applicable law. | D04 permits funding recommendations only when consistent with applicable law. Proposed handling: obtain qualified compliance review before commitment where a credible legal barrier is unresolved. Foreign location or lack of U.S. charitable recognition may require additional diligence, not automatic ineligibility; general-support entity restrictions must be checked separately. | D04. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Counterfactuality failure: the work would happen anyway, or another funder is ready to fund it. | Marginal impact is lower; a referral may be better. | Proposed heuristic: compare what happens without this grant, including timing, scale, quality and displacement of other funding. Verify the realistic alternatives proportionately. Another interested funder does not establish zero additionality; weigh the grant's marginal contribution in the combined case, without a Tier 3 veto. | P04, P10, P12. |

| **Red flag** | **Why it matters** | **Proposed response and source limits** | **Source anchor** |
|---|---|---|---|
| Evidence is anecdote or self-report without a baseline. | You cannot distinguish outcome from selection. | Proposed heuristic: treat anecdotes and self-reports as limited evidence, not proof of causal impact. Seek corroboration, a comparison or sensitivity analysis when the central claim could change the decision. Adjust confidence to the available evidence; lack of a baseline alone is not an automatic tier-based rejection. | D01, P04. |

| Red flag | Why it matters | Proposed response and source limits | Source anchor |
|---|---|---|---|
| Domain-specific red flag: dual-use research, pathogen enhancement, model-weight security, or jailbreak tooling. | Direct catastrophic or misuse risk. | Proposed heuristic: use qualified specialist review to assess the specific activity, exposure and safeguards. A dual-use label alone does not establish unacceptable risk or ineligibility. Pause support for relevant activities while credible severe-harm or legal concerns remain unresolved; D04 does not publish this list as automatic exclusions. | D04, P08. |

**Step 2**

Table 35: Decisive-evidence tests are chosen for their ability to change the decision, not for their completeness. This implements P04's habit and D01's confidence-plus-support requirement.

| Claim to test | Decisive evidence | Efficient method | Update rule |
|---|---|---|---|
| Theory of change is causally plausible. | A concrete mechanism, a load-bearing assumption, and a falsifier. | Ask the applicant to walk through the causal chain and name the assumption most likely to fail. | If no mechanism or no falsifier, decline or refer. If the mechanism is clear, proceed. |

| **Claim to test** | **Decisive evidence** | **Efficient method** | **Update rule** |
|---|---|---|---|
| Track record predicts success on this project. | Independent outcome data from prior comparable work, not a list of outputs. | Reference calls with people who funded or worked with the team; check prior grant reports. | If prior outcomes are weak and no strong advisors are present, decline. If strong, reduce project-level diligence. |
| Strategic judgment is strong. | The applicant can prioritize risks, state what would change their mind, and reason under uncertainty. | Use a live case question or counterfactual; ask them to red-team their own proposal. | If they cannot, decline for Tier 2 and above. If they can, proceed. |
| Project risks are manageable. | Named failure modes, severity, likelihood, mitigation, and no unacceptable residual risk. | Risk-register review; specialist consultation for dual-use or security issues. | If catastrophic residual risk is unmitigated, stop. If manageable, set milestones. |
| Cost-effectiveness is plausible. | Reproducible BOTEC with unit costs, a denominator, a counterfactual, and sensitivity. | Rebuild the BOTEC; vary key parameters; compare to reference classes. | If the estimate is fragile or double-counted, reduce the grant or decline. If robust, proceed. |
| Financial and governance controls are sound. | Budget versus actuals, audit, runway, related-party disclosure, conflict policy. | Request documents; ask for the last grant report; check financial references. | If controls are weak, stop unless a fiscal sponsor or staged funding mitigates. |
| Founder or team can execute. | References from close collaborators, evidence of shipping, co-founder or hiring plan. | Back-channel references; ask what they have shipped, not what they plan. | If the founder is not exceptional and the problem is not important, stop. Otherwise stage. |

| Claim to test | Decisive evidence | Efficient method | Update rule |
|---|---|---|---|
| Counterfactuality is real. | Evidence that the grant changes whether the work happens. | Ask who else is funding; ask what happens if CG says no; check co-funding. | If fully substitutable, refer. If marginal, fund at a lower amount or stage. |
| Scale and absorption are realistic. | Hiring plan, management capacity, timeline, dependencies, and WBS. | Review the WBS and hiring pipeline; ask for a milestone plan. | If absorption is implausible, fund smaller or refer. If plausible, proceed. |

**Step 3**

Pre-commit the stop rule before investigation. The stop rule has five forms: fatal stop, budget stop, time stop, referral stop, and ambiguity stop. Write the chosen stop rule into the grant memo so that it is not reconstructed after the fact.

- Fatal stop: any legal, safety, integrity, or catastrophic-residual-risk red flag. No amount of offsetting benefit justifies proceeding without mitigation.
- Budget stop: stop when the expected value of additional information is less than the cost of obtaining it, including grantee and staff time.
- Time stop: for Tier 0, one working day for a decision or clarification. For Tier 1, one to four weeks. For Tier 2, two to six weeks. For Tier 3, one to three months. If the decisive evidence is not available by the pre-committed date, decide with what you have or decline.
- Referral stop: if the grant is below CG's comparative-advantage threshold and a specialized funder is better positioned, refer rather than investigate. D04 names SFF, LTFF, Founders Pledge, AI Risk Mitigation Fund, and Longview. P10's Transformative AI Fund and P13's BlueDot Rapid Grants cover smaller tiers.
- Ambiguity stop: if evidence is mixed but the downside is tolerable, stage the grant, set milestones, and convert the remaining uncertainty into a future decision. Do not fund the full amount just to avoid a hard call.

$$\text{EVSI} = \mathbb{E}_y\Big[\max_a \mathbb{E}[U(a,\theta) \mid y]\Big] - \max_a \mathbb{E}[U(a,\theta)]$$

$$\text{EVSI} > c_{\text{investigation}} + c_{\text{grantee time}}$$

Here a is an available decision, $\theta$ is the uncertain state, and y is the possible result of the investigation. Average over possible findings and allow the decision to change after seeing them. Compare that expected improvement with the costs of obtaining the information. Without the decision change and the expectation over possible findings, the previous expression was not a complete ex ante value-of-information calculation.

**Step 4**

Over-diligence failure mode. Due diligence is two-sided. P04 says the goal is the most decision-relevant information as time-efficiently as possible while avoiding unnecessary work for the organizations being evaluated. D04 says CG is not planning to provide feedback for most expressions of interest because of capacity.

*A non-trivial point: a stop rule must be able to stop because information is not worth its cost, not only because evidence is negative. Otherwise the investigator defaults to more diligence, which is safe for the individual but costly for the portfolio.*

Dealflow cost: excessive diligence on small grants filters out exactly the early-stage founders and small organizations that P10, P11, P12, P13 and P14 argue are the field's bottleneck.

## Sub-Question 4: Worked Drill: $10m Multi-Year Grant, $5k Rapid Grant, and $200k-$2m First-Time Grant

Apply the tiers to three cases. For each, state the tier, the decisive evidence, the stop rule, and two defensible answers. The point is not to memorize thresholds but to practice moving between evidence thresholds.

**Step 1**

Case A: a $10m multi-year grant. Tier 3. Evidence threshold: independent outcome evidence, leadership references, external red-team, financial audit or equivalent, legal due diligence, co-funder coordination, and a staged agreement with milestones.

- Decisive tests: What would change our mind about the team? What evidence shows the organization can absorb $10m without governance failure? What is the catastrophic downside if the strategy is wrong?
- Stop rule: fatal stop on legal, safety, integrity, or unmitigated catastrophic risk. Stop if two load-bearing claims remain unsupported after the investigation budget. Stage if uncertainty is tolerable.
- Two defensible answers. Answer A: fund because track record and independent evidence are strong, and the cost of delay is high. Answer B: decline or stage because the theory of change is not falsifiable and the

downside is irreversible. The crux is whether the load-bearing uncertainty can be resolved by milestones. If yes, stage. If no, decline.

**Step 2**

Case B: a $5k rapid grant. Tier 0. Evidence threshold: concrete work in motion, specific cost bottleneck, reason the work matters, eligibility, and no sanctions or fraud red flag.

- Decisive tests: Has the applicant already started? Is the cost the bottleneck? Would the work stop without this grant? Source anchors: P13, P14, D03.
- Stop rule: if there is no artifact or no bottleneck, decline. If ambiguous, ask one clarification question. Decide within five working days. Pay upfront where possible.
- Two defensible answers. Answer A: fund quickly to preserve dealflow and encourage small projects. Answer B: raise the bar as the program matures, as P13 describes, and decline anything without traction. The crux is whether selection effects or speed effects dominate at this stage.

**Step 3**

Case C: a $200k-$2m first-time grant. Tier 2. Evidence threshold: D04's six criteria, with the EOI focusing on theory of change, track record, and strategic judgment, and the full investigation adding project risks, cost-effectiveness, and scale.

- Decisive tests: Can the applicant name what would change their mind? Is the budget justified by cost drivers and work packages? Is there a plausible counterfactual?
- Stop rule: if below $200k per year, refer to other funders unless the proposal has been rejected elsewhere. If above threshold, fund one to two years and renew after evidence. Top up only after evidence of success.
- Two defensible answers. Answer A: fund as a one-year pilot because the team is strong and the theory of change is plausible. Answer B: decline because the track record is not in this exact area and the project is not yet above the bar. The crux is whether strong general capability plus staged funding is enough to substitute for domain-specific track record.

## Sub-Question 5: Calibrated Interview Answer and Rehearsal

Produce a concise answer that makes assumptions and limitations visible. It should show craft, not credentials.

**Step 1**

I treat due diligence as an information-acquisition problem under a budget. I set the depth by three axes: downside severity, uncertainty, and ticket size. Ticket size sets approval and expected value of information, but it is a weak proxy for diligence; a small dual-use grant can need more scrutiny than a large renewal.

I follow D04's stated emphasis: at EOI I focus primarily on theory of change, track record, and strategic judgment; during full review I examine all six considerations holistically, including evidence of promise where a directly matching track record is absent. I pre-commit a stop rule: fatal red flags stop immediately; otherwise I stop when the expected value of additional information is less than the cost, when the time budget expires, or when the grant is below our comparative advantage and should be referred.

For small grants, I use light, fast evidence thresholds and pay upfront where possible. For large grants, I require independent outcome evidence, leadership references, legal and financial due diligence, and staged funding. I state confidence and what would change my mind in the write-up. Limitations: CG's internal risk tolerance and red lines are not fully public, and D04's thresholds are program-specific, so I would ask an interviewer what the actual current bar is.

**Step 2**

Table 36: The rehearsal makes the proportionality reasoning explicit and avoids inventing credentials or experience.

| **Likely follow-up** | **Calibrated response** | **What not to do** |
|---|---|---|
| How do you decide when to stop investigating? | Pre-commit a time and information budget; stop when EVI is below cost or a fatal red flag appears; state the stop rule in the memo. | Do not say you investigate until certainty; that destroys dealflow and is not proportionate. |
| How much evidence is enough for a $10m grant? | Independent outcome evidence, reference checks, external red-team, financial and legal review, and staged milestones. The threshold is the decision, not a document count. | Do not rely on a single impressive founder reference or a strong narrative. |

| Likely follow-up | Calibrated response | What not to do |
|---|---|---|
| Would you fund a $5k grant with almost no evidence? | I would fund concrete work in motion with a specific cost bottleneck and no integrity red flags. I would not fund a plan that has not started. | Do not apply large-grant evidence standards to rapid grants. |
| What if the project is high-risk but potentially high-reward? | Risk severity can dominate size. I would bump the tier, require specialist review, and either stop or stage. I would not let expected value alone override catastrophic downside. | Do not treat high expected value as a reason to skip downside analysis. |
| How do you handle uncertainty you cannot resolve? | Use staged funding and milestones; convert uncertainty into a future decision. If unresolved uncertainty concerns a fatal risk, stop. | Do not fund the full amount just to avoid a hard call. |

**Step 3**

- Source-supported: the role brief's proportionate-due-diligence requirement; D04's criteria and scale ranges; P04's evidence-seeking habit; L01's budget method; D01's reasoning transparency.
- Inference: the specific tier boundaries, the bump-one-tier rule for high downside, and the five stop-rule types are proposed craft, not CG statements.
- Open questions: What is CG's current internal risk tolerance and red-line list? How do actual grantmakers weigh track record versus theory of change at the margin? How has the bar changed after the 2026 scale-up? What counterfactuality and credit-allocation conventions does CG use for co-funded grants?
- Interview use: state the method, then ask the interviewer for the actual current thresholds. This converts an open question into a demonstration of calibration.

# Learn grant-memo writing and WBS, CBS, OBS and RACI budget construction.

## Question

Build the two artifacts a grantmaker actually produces: a **grant memo** that a decision-maker can update on, and a **budget** built by decomposition rather than by guessing a total. This solution provides a memo template, a hypothetical worked cost table, and a RACI example. The completed illustration below shows how to state a decision when key evidence is missing. The cost-effectiveness verdict and multi-parameter sensitivity exercise remain open until the learner supplies justified assumptions; they are not completed results.

The non-trivial requirement is not learning four acronyms. It is seeing that the four structures are **four different cuts through the same object**, each exposing a different class of error: the WBS exposes work that is not attached to a deliverable, the CBS exposes cost that is not attached to a single driver, the OBS exposes deliverables that have no owner, and the RACI matrix exposes accountability that is diffused or overloaded. A budget that passes all four cuts is one whose failure modes are visible before the money moves.

## Background

Source status matters here more than usual. L01 ("How to ask for money") is a short personal note, written in Polish, that the author himself describes as an informal account of standard project-management practice, reconstructed from reading done long ago, with the explicit caveat that he is not certain he has not left something out. It also asserts that this structure is uncommon in EA organizations, where a single very capable person often keeps every work package in their head. Every claim from L01 should therefore be tagged as **proposal or method**, not as a description of how Coefficient Giving writes memos.

Three other sources supply the surrounding norms. D01 ("Reasoning Transparency") gives the reader-serving requirements for any analysis. L02 ("What Good Writing Looks Like") gives a general account of reader affordances and a six-move argument structure, and its worked examples explicitly contain invented premises. D04 (the AI Governance RFP) gives the assessment criteria CG itself publishes, including theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale, plus a default first-grant band of $200k-$2M/yr over one to two years. P04 illustrates constrained grantmaker time through separate accounts from Trevor Levin, Chris Bakerlee, and Abbey Chaver; it does not supply a universal two-hour memo or investigation budget.

D02 supplies the strongest indirect evidence that CG already values exactly this decomposition: CG built an internal **Grant Copilot** to draft memo write-ups and a **BOTEC generator** to produce back-of-the-envelope calculations

from a grant idea, and calls a back-of-the-envelope calculation a thing CG loves because it calibrates and quantifies scope and cost-effectiveness. That is an inference about institutional practice, not a statement of policy, and it should be labelled as such.

## Sub-Question 1: Grant-memo architecture: the six moves, the three requirements, and proportionate diligence

Design the memo as a decision instrument for one named reader, then decide how much evidence to gather before writing it. The memo template must map every section to the source that requires it and to the failure mode it prevents. The diligence decision must be treated as an expected-value-of-information problem rather than as "more diligence is better".

### Step 1

D01 and L02 are not two rubrics to satisfy separately; they answer two different questions about the same text. D01 asks *"how should I discount this claim?"* and L02 asks *"what do I need in order to take the next inferential step?"* A grant memo has to do both, because its reader is deciding whether to move money and its writer is asking the reader to follow a chain from evidence to recommendation.

The non-obvious move is recognizing that **the falsifier field does triple duty**. L02 asks the writer to show how they could be moved; D04 lists project risks as an assessment criterion with a warning that listing every possible risk is worse than prioritizing; and P04 describes the habit of asking early what piece of evidence would most likely change your mind and then seeking it out. The same object plays all three roles, which is why a memo without it is not merely incomplete but unfalsifiable.

Table 37: A memo map that ties each section to the source that requires it and to the error it prevents.

| **Memo section** | **Required by** | **What it is for** | **Failure mode if omitted** |
|---|---|---|---|
| Linked summary of key takeaways | D01 | lets the reader decide whether to read on, and where to jump | the reader either reads everything or nothing |
| Importance ranking of considerations | D01 | makes the holistic judgment inspectable | an unfalsifiable holistic verdict |

| Memo section | Required by | What it is for | Failure mode if omitted |
|---|---|---|---|
| Confidence plus support kind per major claim | D01 | lets the reader discount each claim differently | every claim looks as strong as the strongest one |
| What would change our mind | L02 move 3; D04 risks | turns the recommendation into a claim rather than a preference | the memo cannot be argued with, only overruled |
| Evidence and provenance | L02 move 4; D01 | tells the reader how far to trust the check | diligence cost is invisible and unreproducible |
| Counterfactuality | P02 and P05 ROI method | separates our effect from the background | impact double-counted against co-funders |
| Budget by cost pool with a driver each | L01 | makes the money legible in one glance | the reviewer must audit line by line |

**Exercise:** Supply the named input explicitly before running this code. No result from this exercise has been calculated here.

```
WEIGHTS = {
    'linked_summary': 1.0,
    'importance_ranking': 1.0,
    'confidence_and_support': 1.0,
    'meet_reader': 0.5,
    'claim_and_reasoning': 1.0,
    'falsifier': 1.0,
    'evidence_provenance': 1.0,
    'criterion_applied': 1.0,
    'recommended_update': 1.0,
    'cost_drivers_justified': 1.0,
    'counterfactuality': 1.0,
}

SUPPORT_KINDS = {
    'own_analysis', 'expert_qualified', 'expert_unassessable',
    'skim_qualified',
    'skim_weak', 'verifiable_sourced', 'verifiable_unsourced',
    'reference_class',
    'general_intuition', 'undocumented_conversations',
}

```

```
def audit_memo(memo, weights):
    gaps, score = [], 0.0
    total = sum(weights.values())
    for field, w in weights.items():
        value = memo.get(field)
        filled = bool(value.strip()) if isinstance(value, str) else bool(value)
        if filled:
            score += w
        else:
            gaps.append(field)
    if memo.get('confidence_and_support') and memo.get('support_kind') not in SUPPORT_KINDS:
        gaps.append('support_kind_unlabelled')
    return gaps, round(score / total, 3)

memo = input_state['memo']
gaps, completeness = audit_memo(memo, WEIGHTS)

template = [
    ('summary_of_key_takeaways', 'one line each, each linked to the section that argues it'),
    ('question_and_recommendation', 'amount, term, instrument, conditions precedent'),
    ('theory_of_change', 'activity to output to outcome to impact, with the assumed linkage made explicit'),
    ('what_would_change_our_mind', 'the crux; the observation that would flip the recommendation'),
    ('evidence_and_provenance', 'what was checked, how deeply, what could not be checked, and why'),
    ('counterfactuality', 'why this happens with us and not without us'),
    ('risks_and_failure_modes', 'ranked by scale and likelihood, not enumerated'),
    ('budget_by_cost_pool', 'pool, single cost driver, quantity, unit cost, justification kind'),
    ('kpis_and_measurement', 'what will be counted, when, and what the count would trigger'),
    ('open_questions_and_confidence', 'per major claim: confidence level and support kind'),
]

output_state['memo_gaps'] = gaps
output_state['memo_completeness'] = completeness
output_state['memo_template'] = template
```

Note the deliberate asymmetry: **a memo can state a confidence and still fail the audit**, because D01 requires both the confidence and the kind of support behind it. An unlabelled support kind is scored as missing. This is where

L01's allowance for vibes-based measures collides with D01. L01 permits a KPI that is literally "on a scale of one to seven, how much would Dustin Moskovitz want this project to continue", but D01 would force that measure to be labelled as an ordinal proxy with a stated reason for having no better one.

**Step 2**

P04 supplies three distinct examples. Trevor Levin describes spending two hours finishing one write-up on a representative Tuesday. Chris Bakerlee reports investigating 6–12 grants at different levels of attention. Abbey Chaver describes dividing her time roughly equally among meetings, research, and investigations or memos. These observations come from different people and do not establish a two-hour end-to-end grant investigation or a universal writing budget. Levin's related advice is to seek decision-relevant evidence efficiently while avoiding unnecessary work for the organizations being evaluated. D04 adds that cost-effectiveness assessment counts additional costs a project imposes on others, such as the time required from evaluators or collaborators. So "how much checking is proportionate" is not a matter of taste; it is a decision problem with an answer.

$$\mathrm{EVOI}(\mathrm{check}) = \Big[P(\mathrm{pass})\max\big(0,\ \mathbb{E}[V \mid \mathrm{pass}]\big) + P(\mathrm{fail})\cdot 0\Big] - \max\big(0,\ \mathbb{E}[V]\big) - c$$

The quantity being computed is the value of the **decision rule that funds only on a pass**, not the value of the information itself. That distinction matters: a check can be highly informative and still worth skipping if the prior decision does not change.

```
import numpy as np

def ev_of_policy(p, G, B):
    # value of the best decision available on a belief p that the
    grant is above the bar
    return max(0.0, p * G + (1.0 - p) * B)

def evoi(p, G, B, sens, spec, cost):
    base = ev_of_policy(p, G, B)
    p_pass = p * sens + (1.0 - p) * (1.0 - spec)
    p_fail = 1.0 - p_pass
    post_pass = (p * sens) / p_pass if p_pass > 0 else 0.0
    post_fail = (p * (1.0 - sens)) / p_fail if p_fail > 0 else 0.0
    funded = p_pass * ev_of_policy(post_pass, G, B)
    return funded - base - cost, post_pass, post_fail

p = input_state['decision_parameters']['p_prior']
G = input_state['decision_parameters']['value_if_above_bar']
```

```
B = input_state['decision_parameters']['value_if_below_bar']

rows = []
for c in input_state['checks']:
    net, post_pass, post_fail = evoi(p, G, B, c['sensitivity'],
    c['specificity'], c['cost'])
    rows.append({'check': c['name'],
                 'net_evoi': round(net, 1),
                 'posterior_if_pass': round(post_pass, 3),
                 'posterior_if_fail': round(post_fail, 3),
                 'break_even_cost': round(net + c['cost'], 1)})
rows.sort(key=lambda r: -r['net_evoi'])

priors = np.linspace(0.01, 0.99, 99)
curves = {c['name']: [evoi(x, G, B, c['sensitivity'],
    c['specificity'], c['cost'])[0] for x in priors]
          for c in input_state['checks']}

output_state['evoi_table'] = rows
output_state['evoi_curves'] = {'priors': priors.tolist(), 'curves':
    curves}
output_state['best_check'] = rows[0]['check'] if rows else None
output_state['break_even_cost'] = rows[0]['break_even_cost'] if rows
    else None
```

Two caveats belong in the memo itself. First, the value units must be commensurable for the subtraction to mean anything: converting a grantmaker-hour into CG-dollars is a judgment, and D01 would require it to be labelled as such. Second, **checks are usually correlated**. If the second expert call is drawn from the same network as the first, its effective sensitivity is far below its standalone value, and the model will overstate the case for running it.

## Sub-Question 2: Work breakdown structure: from theory of change to packages, schedule, milestones and tiers

Decompose a theory of change into work packages, then into activities with durations and dependencies, derive a schedule and milestones, and finally express the work as a nested ladder of tiers in which each increment is a coherent bundle that moves a named KPI.

### Step 1

L01's sequence is: the theory of change is decomposed into work packages; a work package is a bundle of obligations that contributes to a deliverable; each deliverable contributes to a goal in the theory of change; each package should be measured against a KPI even if the measure is weak; packages break into activities with durations and dependencies, which produce a schedule; mile-

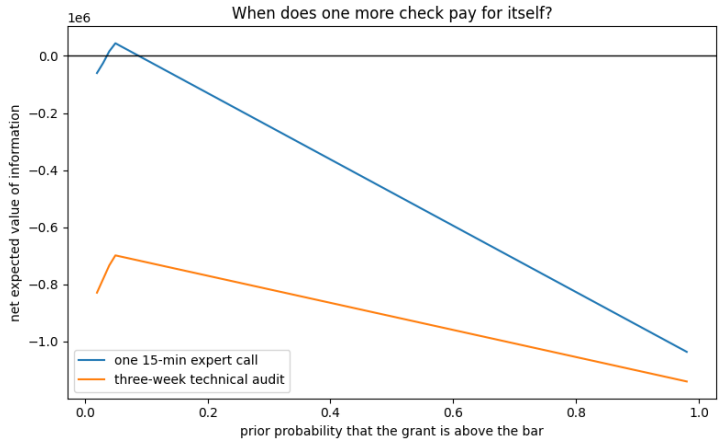


Figure 15: Graphical analysis for Step 2.

stones follow from delivering significant deliverables; and once you have the work structure and the cost structure you can present tiers.

The non-obvious consequence is that **the schedule sort is a real computation, not bookkeeping**. With finish-to-start dependencies, the project duration is the longest path through the activity graph, and only activities on that path have the property that their slippage moves the milestone. Every other activity has float. This is what turns L01's "estimate durations and dependencies" from a spreadsheet exercise into a decision: adding capacity to a package with float buys nothing, and a package on the critical path with one owner is the schedule's single point of failure.

```
def topo_and_schedule(activities):
    by_id = {a['id']: dict(a, start=0.0, finish=0.0) for a in
    activities}
    order, in_progress = [], set()

    def visit(nid):
        if nid in order:
            return
        if nid in in_progress:
            raise ValueError('dependency cycle at ' + nid)
        in_progress.add(nid)
        for d in by_id[nid]['deps']:
            visit(d)
        in_progress.discard(nid)
        order.append(nid)

    for nid in list(by_id):
        visit(nid)
```

```
    for nid in order:
        a = by_id[nid]
        a['start'] = max([by_id[d]['finish'] for d in a['deps']], default=0.0)
        a['finish'] = a['start'] + a['weeks']
    return by_id, order

def trace_critical_path(by_id):
    node = max(by_id.values(), key=lambda a: a['finish'])
    path = []
    while node is not None:
        path.append(node['id'])
        drivers = [by_id[d] for d in node['deps'] if abs(by_id[d]['finish'] - node['start']) < 1e-9]
        node = drivers[0] if drivers else None
    return list(reversed(path))

spec = input_state['project_spec']
all_activities = [a for wp in spec['work_packages'] for a in wp['activities']]
by_id, _ = topo_and_schedule(all_activities)
cp = trace_critical_path(by_id)
total_weeks = max(a['finish'] for a in by_id.values())

wbs, milestones = [], []
for wp in spec['work_packages']:
    acts = [by_id[a['id']] for a in wp['activities']]
    wbs.append({
        'work_package': wp['id'],
        'deliverable': wp['deliverable'],
        'kpi': wp['kpi'],
        'cost_driver': wp['cost_driver'],
        'activities': [{'id': a['id'], 'name': a['name'], 'weeks': a['weeks'],
                        'start': a['start'], 'finish': a['finish']} for a in acts],
        'package_finish_week': max(a['finish'] for a in acts),
        'on_critical_path': any(a['id'] in cp for a in acts),
    })
    milestones.append({'milestone': wp['deliverable'], 'week': max(a['finish'] for a in acts)})

output_state['wbs'] = wbs
output_state['schedule'] = sorted([by_id[a['id']] for a in all_activities], key=lambda a: a['start'])
output_state['critical_path'] = cp
output_state['total_duration_weeks'] = total_weeks
output_state['milestones'] = sorted(milestones, key=lambda m: m['week'])
```

L01's own observation is that this discipline is uncommon in EA organizations, where a few very capable people hold the entire set of work packages in their heads and can assign them to specific people and notice when important work is not being done. That works at small scale and fails at larger or longer projects, which is precisely the condition CG is now in. But this is L01's claim about EA organizations generally, not a verified claim about CG, and it should be presented as a risk hypothesis rather than a fact.

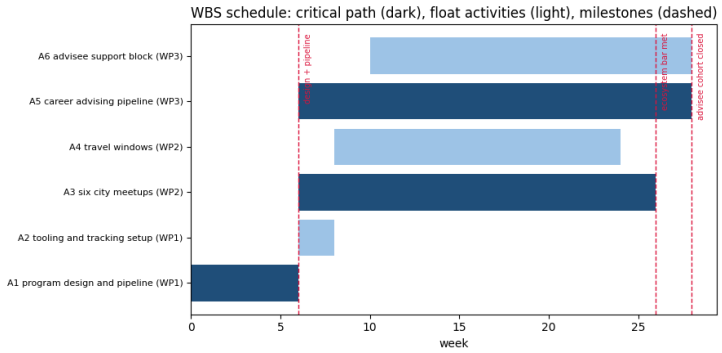


Figure 16: Graphical analysis for Step 1.

### Step 2

L01's last constructive step is that a WBS plus a CBS lets you offer several tiers: bundles of work items that contribute to a specific goal, depend on specific cost drivers, and ideally produce a specific increase in a KPI. The non-obvious reading is that **a tier is not a discount version of the same project**. It is a package a funder can actually buy, which imposes three structural constraints the WBS must satisfy.

- Nesting: a higher tier is a superset of the lower one, so the funder is buying an increment rather than a different project.
- KPI linkage: every activity in the increment has a named KPI link, otherwise the funder is buying activity rather than outcome.
- A named funder bar: each tier names the funder whose threshold it is designed to clear. A tier ladder with no named buyer is a wish list.

```
def validate_tiers(tiers, wbs):
    activity_ids = {a['id'] for wp in wbs for a in wp['activities']}
    flags, prev = [], set()
    for name, spec in tiers.items():
        ids = set(spec['activities'])
        unknown = ids - activity_ids
        if unknown:
```

```
            flags.append({'tier': name, 'issue':
    'unknown_activities', 'detail': sorted(unknown)})
        if not prev.issubset(ids):
            flags.append({'tier': name, 'issue':
    'not_a_superset_of_the_previous_tier'})
        increment = ids - prev
        unlinked = sorted(a for a in increment if a not in
    spec.get('kpi_linkage', {}))
        if unlinked:
            flags.append({'tier': name, 'issue':
    'increment_without_kpi_linkage', 'detail': unlinked})
        if not spec.get('funder_bar'):
            flags.append({'tier': name, 'issue':
    'no_funder_bar_named'})
        prev = ids
    return flags

tiers = input_state['tiers']
flags = validate_tiers(tiers, input_state['wbs'])
output_state['tier_validation_flags'] = flags
output_state['tier_ladder_ok'] = len(flags) == 0
output_state['tier_kpi_deltas'] = {t: s.get('kpi_delta') for t, s in
    tiers.items()}
```

One caution that L01 does not draw but its own logic implies: **the KPI attached to the top tier's increment is usually the weakest-measured one**, because the increment is the part of the project that does not exist yet. That is where L01's "vibes" allowance does its real work, and where D01's confidence-and-support requirement bites hardest.

## Sub-Question 3: Cost breakdown structure: cost pools, single drivers, coverage, and justification

Convert the WBS into a cost structure organized by cost driver rather than by deliverable, check that every work package has cost and every pool has work, attach a justification kind to each pool, and adjust the result for attribution.

### Step 1

L01 is explicit that the cost breakdown structure takes a different axis from the WBS. The WBS is organized by deliverables; the CBS is organized by cost categories chosen so that each has a single cost driver, and such categories are called cost pools. A resource-based CBS, where the axis is the type of resource, is possible but less informative.

The non-obvious reason for the single-driver rule is stated only briefly in L01: if you watch total cost, you learn too late that you are about to run out of money, whereas if you watch cost drivers you can see one of them starting to rise, such

as rent per square metre, and adapt your plan in time. **A pool with two drivers has no early-warning signal**, because a rise cannot be attributed to a driver. This is why the check is structural rather than cosmetic.

```
def build_cbs(cost_items, wbs):
    pools = {}
    for item in cost_items:
        pool = pools.setdefault(item['cost_pool'], {
            'cost_driver': item['cost_driver'],
            'justification': item['justification'],
            'lines': [],
            'total': 0.0,
        })
        if pool['cost_driver'] != item['cost_driver']:
            pool['mixed_driver'] = True
        if pool['justification'] != item['justification']:
            pool['mixed_justification'] = True
        pool['lines'].append(item)
        pool['total'] += item['quantity'] * item['unit_cost']

    total = sum(p['total'] for p in pools.values())
    for name, p in pools.items():
        p['total'] = round(p['total'], 2)
        p['share'] = round(p['total'] / total, 3) if total else 0.0

    wp_ids = {wp['work_package'] for wp in wbs}
    item_wps = {i['work_package'] for i in cost_items}
    coverage_gaps = {
        'work_packages_without_cost': sorted(wp_ids - item_wps),
        'cost_items_pointing_at_unknown_package': sorted(item_wps -
    wp_ids),
        'pools_with_mixed_driver': sorted(n for n, p in pools.items()
    if p.get('mixed_driver')),
    }

    ranked = sorted(pools.items(), key=lambda kv: -kv[1]['total'])
    return pools, round(total, 2), coverage_gaps, [(n, p['share'])
    for n, p in ranked]

pools, total, gaps, ranking = build_cbs(input_state['cost_items'],
    input_state['wbs'])
output_state['cost_pools'] = pools
output_state['budget_total'] = total
output_state['coverage_gaps'] = gaps
output_state['cost_driver_ranking'] = ranking
```

The coverage check is the part a reviewer should look at first. An empty row means a work package with a deliverable and no money behind it, which is exactly the class of error L01 says the decomposition exists to catch. An empty

column means a cost pool feeding no deliverable, which is where budgets quietly acquire overhead that no one owns.

Table 38: Four cuts through one project; each cut is responsible for a different class of error.

| Structure | Axis | Error class it exposes | Source |
|---|---|---|---|
| WBS | deliverables | work with no deliverable, or a deliverable with no work | L01 |
| CBS | cost drivers, one per cost pool | cost that cannot be attributed to a driver, so no early warning | L01 |
| OBS | roles and functions | a category with no owner, or a role with no category | L01 |
| RAM / RACI | roles by categories | accountability that is diffused or overloaded | L01 |

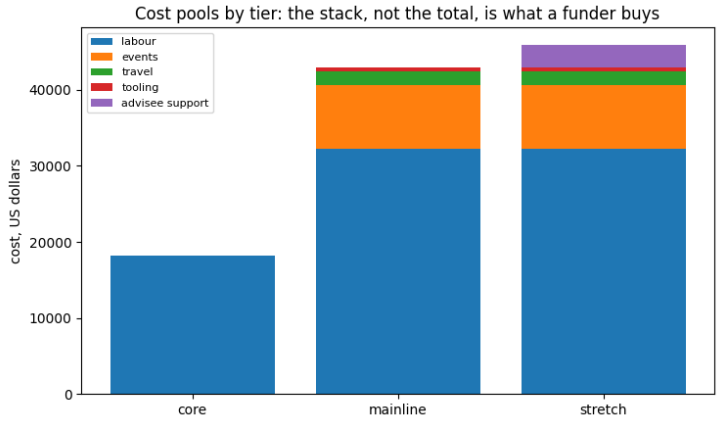


Figure 17: Graphical analysis for Step 1.

### Step 2

L01 gives the reviewer's job description precisely. The person checking the budget should only have to see three things: that the cost drivers really are

what the applicant says they are, so the money will be spent roughly as expected unless a driver changes; that the cost categories correspond to work categories; and that the work categories contribute to the goals in the theory of change. The last of these is the only one that needs evidence, and L01 names three admissible forms of proof: prior spending that produced the effect, a reference class of someone in similar circumstances spending money on the same thing and getting the KPI, or a stated mechanism by which doing the thing raises the KPI.

The mechanism proof is the weakest and should be labelled as such. L01's example is that researchers in one room iterate faster, which is a plausible mechanism with no measurement behind it. Under D01 that claim needs a confidence and a support kind, and under D04 it needs to survive the project-risks criterion.

There is a fourth step L01 mentions in passing and that deserves elevation: the proof of the link between work and goal should be adjusted for attribution. The adjustment is the counterfactuality problem, and it appears everywhere else in the corpus. P02 reports a mean modelled social return on investment of roughly 5,500x for effective careers grants, weighted by funding, and immediately flags the estimates as early and uncertain. P05 reports an average weighted return on investment of roughly 5x to 6x for its effective giving portfolio after quality and counterfactuality adjustments. Those two numbers differ by three orders of magnitude and are not describing the same quantity, which is itself the lesson: **the counterfactuality multiplier, not the raw cost, is what makes a cost-per-outcome number comparable across grants**.

$$\text{cost per attributable unit} = \frac{\text{pooled cost}}{\text{units} \times \text{counterfactuality}}$$

```
JUSTIFICATION_KINDS = {'prior_spend', 'reference_class', 'mechanism',
    'market_price', 'funder_benchmark'}

def attributable_cost_per_unit(cost, units, counterfactuality):
    attributable = units * counterfactuality
    return cost / attributable if attributable > 0 else float('inf')

pools = input_state['cost_pools']
total = input_state['budget_total']

unjustified = [name for name, p in pools.items() if
    p['justification'] not in JUSTIFICATION_KINDS]
unjustified_share = sum(pools[n]['total'] for n in unjustified) /
    total if total else 0.0

units = input_state['kpi_assumption']['units']
cf = input_state['kpi_assumption']['counterfactuality']
base = attributable_cost_per_unit(total, units, cf)
```

```
16
17 grid = {round(x, 2): round(attributable_cost_per_unit(total, units,
       x), 1)
18         for x in [0.2, 0.4, 0.6, 0.8, 1.0]}
19
20 output_state['unjustified_pools'] = sorted(unjustified)
21 output_state['unjustified_share'] = round(unjustified_share, 3)
22 output_state['cost_per_attributable_unit'] = round(base, 1)
23 output_state['counterfactuality_sensitivity'] = grid
24 output_state['largest_cost_driver'] = max(pools.items(), key=lambda
       kv: kv[1]['total'])[0] if pools else None
```

The honest caveats. Resource-based costing is the convention in practice even though L01 finds it less informative, so a memo that uses cost pools may look unfamiliar to a reviewer. The single-driver rule breaks down for shared fixed costs, which have no natural driver and must be allocated by convention. And L01 itself notes that the whole framework is standard practice that it is reconstructing from memory, so the memo should not present the structure as a house standard it cannot verify. S10 supplies the complementary failure mode from the other side of the table: funding based on cost rather than value delivered, funding games of chicken between funders that produce glacial response times, and unclear credit allocation with double-counted impact. A cost pool table with a counterfactuality column is precisely an instrument against all three.

## Sub-Question 4: OBS, control accounts and RACI: putting an owner on every deliverable

Map WBS categories to roles, form control accounts at the intersections, build the RACI matrix, and validate it against the failure modes the structure is designed to expose: unowned work, roles without control accounts, and overloaded accountability.

### Step 1

L01 describes the OBS as the structure in which you record which WBS categories are assigned to which role, and the point at the intersection of a WBS category and a role as the control account: the place where ownership of that category lives. A single role can own several categories, a role can own none, and a category can have no owner. L01 calls this a tool for assigning responsibilities, which is a modest description of something stronger: **a work package with a deliverable and no owner will not happen, and no budget line fixes that**.

The RACI refinement adds a validation rule that a plain OBS does not have. Each work package should have exactly one accountable role, at least one responsible role, and any number of consulted and informed roles. Exactly one is

the operative phrase: with two accountable roles, accountability is diffused and neither will act when the milestone slips; with zero, the package is unowned.

```
def build_raci(assignments, work_packages, roles, max_span=2):
    matrix = {(r, w): set() for r in roles for w in work_packages}
    for a in assignments:
        matrix[(a['role'], a['work_package'])].add(a['raci'])

    flags, control_accounts = [], []
    for w in work_packages:
        accountable = [r for r in roles if 'A' in matrix[(r, w)]]
        responsible = [r for r in roles if 'R' in matrix[(r, w)]]
        if len(accountable) != 1:
            flags.append({'work_package': w, 'issue':
    'accountable_count', 'count': len(accountable)})
        if not responsible:
            flags.append({'work_package': w, 'issue':
    'no_responsible_role'})
        if not accountable and not responsible:
            flags.append({'work_package': w, 'issue':
    'unowned_work_package'})
        if len(accountable) == 1:
            control_accounts.append({'work_package': w,
    'accountable_role': accountable[0]})

    for r in roles:
        owned = [w for w in work_packages if 'A' in matrix[(r, w)]]
        if not owned:
            flags.append({'role': r, 'issue':
    'role_without_control_account'})
        elif len(owned) > max_span:
            flags.append({'role': r, 'issue':
    'accountable_span_over_limit',
                          'count': len(owned), 'limit': max_span})

    flat = {r + ' :: ' + w: ''.join(sorted(matrix[(r, w)])) for (r,
    w) in matrix}
    spans = {r: sum(1 for w in work_packages if 'A' in matrix[(r,
    w)]) for r in roles}
    return flat, control_accounts, flags, spans

flat, control_accounts, flags, spans = build_raci(
    input_state['assignments'], input_state['work_packages'],
    input_state['roles'])

output_state['raci_matrix'] = flat
output_state['control_accounts'] = control_accounts
output_state['raci_flags'] = flags
output_state['accountable_span_by_role'] = spans
```

The span rule is the one that matters at CG's scale. Setting the limit at two accountable packages per role is a choice, not a source claim, and the code takes it as a parameter. The reason to have some limit is that a role accountable for many packages is a role that will be the bottleneck on the schedule and the point of failure in the OBS. L01's footnote that a function may have no categories at all reads as an allowance; under a scaling organization it is better read as a flag.

One tempting but unearned analogy: D02 describes CG appointing a Directly Responsible Individual for AI enablement, which is the same concept as the accountable role. The concept match is real, but D02 is about internal tooling adoption, not grantmaking, so drawing the inference that CG runs RACI on grants is not supported by that post. That is an inference, and it should be labelled one.

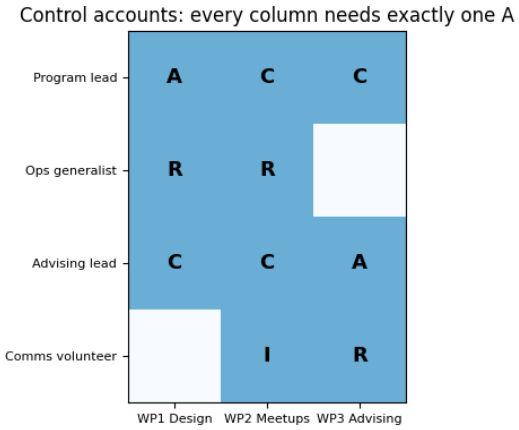


Figure 18: Graphical analysis for Step 1.

### Step 2

The validating artifact is the filled matrix, and it is worth producing even when the answer is already known, because the matrix makes three things visible at once that prose hides.

Table 39: The RACI form of control accounts for the worked example. Read the Accountable column as the schedule's bottleneck list.

| Work package | Deliverable | Accountable | Responsible | Consulted / Informed |
|---|---|---|---|---|
| WP1 | Program design and speaker pipeline | Program lead | Ops generalist | Advising lead (C); Comms volunteer (I) |
| WP2 | Six city meetups with recorded talks | Ops generalist | Ops generalist; Comms volunteer | Program lead (C); Advising lead (C) |
| WP3 | Advising pipeline with 40 active advisees | Advising lead | Comms volunteer | Program lead (C) |

- Diffused accountability: two A's in one column is a package that will stall, because each role can attribute the delay to the other.
- Orphaned capacity: a role with no control account is either underused or attached to work that never made it into the WBS.
- Concentrated risk: one role accountable for many packages is the single point of failure that L01 warns about when a small organization keeps the whole structure in one person's head.

The matrix is a design artifact, not a compliance artifact. Its value in a grant memo is that it lets the reviewer ask a question that a theory of change cannot answer: who specifically will notice if this does not happen. Under D04's track record criterion, that is close to the question of whether the applicant has executed before.

## Sub-Question 5: Worked example: a tiered first-year budget with cost per outcome and a named funder bar

Assemble one project end to end: a first-year national AI-safety field-building program with three tiers, a joint budget across the four structures, an average and a marginal cost per outcome, and an explicit statement of which funder each tier is aimed at and why.

### Step 1

The worked example is deliberately drawn from a documented niche rather than invented from nothing. N03 describes a real national AI-safety field-building

program with a stated theory of change, per-event attendance figures, a BlueDot Rapid Small Grant of $4,800 received in under two hours, unpaid volunteers, and an executive director at half time. D03 describes typical rapid grants as up to $10,000, average decision within a day, nine in ten applicants hearing back within a week. P14 gives the older version of the same program: most grants between $50 and $1,500, higher amounts requiring evidence of traction, and work that is already in motion. S06 states the same program is looking to triple from $50,000 across 76 people to roughly $150,000.

That sets up the funder-fit statement, which is the part a generic template leaves out. D04 says that when funding a project or organization for the first time, the AI Governance RFP typically prefers to begin with $200,000 to $2,000,000 per year over one to two years. P10 gives the small end of the ecosystem: a fund whose grants are typically below $150,000, rarely above $300,000, and not above roughly $1,000,000 for the foreseeable future. So the illustrative stretch tier of $45,880 sits below this AI Governance RFP starting preference, with exceptions; the memo should distinguish this program-specific preference from eligibility. A tier ladder makes the proposed activities, costs, and expected increments visible. Choosing a funder still requires checking the relevant program's remit and criteria. This example has no established renewal amount or renewal commitment; the AI Governance RFP's starting preference should not be transferred to another program.

This is an invented teaching example. All six unit rates, the activity quantities, and the projected transition counts are assumptions. The supplied sources do not establish these particular costs or outcomes. The arithmetic below covers only the six listed cost lines; it is not a verified complete operating budget. Source examples motivate the exercise but do not validate its prices or causal estimates.

```
worked_example = {
    'project': 'National AI-safety field-building program, year 1',
    'funder_fit': {
        'year_one_ask': 'Illustrative mainline amount: $41,080;
    funding route and eligibility unconfirmed',
        'renewal_target': 'No renewal target established; assess the
    relevant program on its own criteria',
        'stated_reason': 'D04 is an AI Governance RFP preference with
    exceptions, not an Effective Careers rule',
    },
    'theory_of_change': {
        'activity': 'meetups, webinars and career advising in a
    country with technical talent but thin field infrastructure',
        'output': 'attendees, advisees, sessions with recorded
    outputs',
        'outcome': 'people enter AI safety or governance training or
    roles',
        'impact': 'capacity added to the field faster than it would
    otherwise have grown',
```

```
        'assumed_linkage': 'contact plus advising converts to transition; this is the load-bearing assumption',
    },
    'work_packages': [
        {'id': 'WP1', 'deliverable': 'program design and speaker pipeline',
         'kpi': 'sessions held with a named, useable output', 'cost_driver': 'organiser-week'},
        {'id': 'WP2', 'deliverable': 'six in-person meetups across five cities',
         'kpi': 'unique first-contact attendees', 'cost_driver': 'attendee-event'},
        {'id': 'WP3', 'deliverable': 'advising pipeline with 40 active advisees',
         'kpi': 'transitions into AI safety work or training', 'cost_driver': 'advisee-month'},
    ],
}

example_cost_items = [
    {'activity': 'A1', 'work_package': 'WP1', 'cost_pool': 'labour', 'cost_driver': 'organiser-week',
     'quantity': 26, 'unit_cost': 700, 'justification': 'illustrative_assumption_unverified'},
    {'activity': 'A2', 'work_package': 'WP1', 'cost_pool': 'tooling', 'cost_driver': 'month',
     'quantity': 12, 'unit_cost': 40, 'justification': 'illustrative_assumption_unverified'},
    {'activity': 'A3', 'work_package': 'WP2', 'cost_pool': 'events', 'cost_driver': 'event',
     'quantity': 6, 'unit_cost': 1400, 'justification': 'illustrative_assumption_unverified'},
    {'activity': 'A4', 'work_package': 'WP2', 'cost_pool': 'travel', 'cost_driver': 'trip',
     'quantity': 10, 'unit_cost': 180, 'justification': 'illustrative_assumption_unverified'},
    {'activity': 'A5', 'work_package': 'WP3', 'cost_pool': 'labour', 'cost_driver': 'organiser-week',
     'quantity': 20, 'unit_cost': 700, 'justification': 'illustrative_assumption_unverified'},
    {'activity': 'A6', 'work_package': 'WP3', 'cost_pool': 'advisee_support', 'cost_driver': 'advisee-month',
     'quantity': 120, 'unit_cost': 25, 'justification': 'illustrative_assumption_unverified'},
]

tiers = {
    'core': {'activities': ['A1', 'A5'],
             'kpi_delta': {'transitions': 3}, 'kpi_linkage': {'A1': 'sessions', 'A5': 'transitions'},
```

```
            'funder_bar': 'Candidate small-grant fund; $32,200 is outside D03's stated typical range; eligibility unconfirmed'},
    'mainline': {'activities': ['A1', 'A2', 'A3', 'A5'],
                 'kpi_delta': {'transitions': 8},
                 'kpi_linkage': {'A1': 'sessions', 'A2': 'sessions', 'A3': 'first-contact attendees', 'A5': 'transitions'},
                 'funder_bar': 'Candidate: EA Funds Transformative AI Fund; amount alone does not establish eligibility'},
    'stretch': {'activities': ['A1', 'A2', 'A3', 'A4', 'A5', 'A6'],
                'kpi_delta': {'transitions': 14},
                'kpi_linkage': {'A1': 'sessions', 'A2': 'sessions', 'A3': 'first-contact attendees',
                                'A4': 'first-contact attendees', 'A5': 'transitions', 'A6': 'transitions'},
                'funder_bar': 'Funding route unconfirmed; do not infer an Effective Careers renewal from D04'},
}

tier_budget = {}
for tier, spec in tiers.items():
    lines = [i for i in example_cost_items if i['activity'] in spec['activities']]
    tier_budget[tier] = {
        'cost': sum(i['quantity'] * i['unit_cost'] for i in lines),
        'kpi_delta': spec['kpi_delta']['transitions'],
        'funder_bar': spec['funder_bar'],
        'n_activities': len(spec['activities']),
    }

output_state['worked_example'] = worked_example
output_state['tiers'] = tiers
output_state['example_cost_items'] = example_cost_items
output_state['tier_budget'] = tier_budget
```

| Activity | Work package | Cost basis | Amount |
|---|---|---|---|
| A1 | WP1: program design and speaker pipeline | 26 organiser-weeks × $700 | $18,200 |
| A2 | WP1: tooling | 12 months × $40 | $480 |
| A3 | WP2: events | 6 events × $1,400 | $8,400 |
| A4 | WP2: travel | 10 trips × $180 | $1,800 |
| A5 | WP3: advising labour | 20 organiser-weeks × $700 | $14,000 |

| Activity | Work package | Cost basis | Amount |
|---|---|---|---|
| A6 | WP3: advisee support | 120 advisee-months × $25 | $3,000 |
| **All listed activities** | | | **$45,880** |

| Tier | Included activities | Total cost | Assumed cumulative transitions | Average cost per transition | Added cost | Added transitions | Marginal cost per added transition |
|---|---|---|---|---|---|---|---|
| Core | A1, A5 | $32,200 | 3 | $10,733.33 | $32,200 | 3 | $10,733.33 |
| Mainline | A1, A2, A3, A5 | $41,080 | 8 | $5,135.00 | $8,880 | 5 | $1,776.00 |
| Stretch | A1, A2, A3, A4, A5, A6 | $45,880 | 14 | $3,277.14 | $4,800 | 6 | $800.00 |

These counts are cumulative, despite the original code's `kpi_delta` label. A1 and A5 remain included in every tier; the marginal columns compare successive totals. The expected transitions are unverified assumptions, so these ratios do not establish cost-effectiveness.

**Illustrative decision memo — no real applicant or funding recommendation**

**Decision:** The funding decision remains open. The modeled one-year mainline program costs $41,080 and assumes eight transitions into AI-safety work or training. Its arithmetic is explicit; its prices, causal effects, execution capacity, and funding fit have not been verified.

**Activity and outcome:** A1 and A5 supply program-design and advising labour; A2 supplies tooling; A3 supplies six events. The proposed mechanism is that contact and advising help participants make transitions they would not otherwise make. Eight transitions is an assumption, not a result from the cited national program.

**Cost and alternatives:** The four included activities sum to $41,080: $18,200 + $14,000 + $480 + $8,400. The core costs $32,200 and assumes three transitions; the stretch costs $45,880 and assumes fourteen. The mainline's raw cost per transition is $5,135. At attributable shares of 0.3 and 0.8, its cost per attributable transition is $17,116.67 and $6,418.75 respectively. These alternatives have different activity sets and unverified outcome assumptions.

**Ownership:** The existing example assigns accountability for design to the program lead, events to the operations generalist, and advising to the advising lead. These are role labels, not confirmed people or evidence of execution capacity.

**Crux and next evidence:** Does the program cause additional transitions at a worthwhile cost? Establish participant counting and attribution, substantiate the unit rates and complete budget, and check who can deliver the activities. Evidence that the same transitions would occur without the program would weaken the case. No measured outcome value or relevant funding threshold is supplied, so a favorable cost-effectiveness verdict is not established.

**Funding route and uncertainty:** The total exceeds D03's typical rapid-grant range. D04's AI Governance preference permits exceptions but is not a rule for Effective Careers. No funder eligibility, renewal, prior rejection, or commitment is assumed. The corpus supports the questions to investigate; the example supplies only hypothetical inputs.

Notice what the example's theory of change makes explicit: the load-bearing assumption is that contact plus advising converts to transition. L02's move three says the memo must show how it could be moved, and here that is concrete. If participants who receive advising are the same people who would have found AI safety work anyway, the counterfactuality multiplier collapses and the cost per attributable transition rises without the cost changing at all.

### Step 2

The average cost per outcome and the marginal cost per outcome are different numbers, and the funder buys the marginal one. This is the single most common error in a tiered budget: a program with an excellent average and a terrible top tier is presented as if the average were the price. The decomposition makes the two separable, because each tier's increment has its own activity set and its own KPI delta.

$$\text{marginal cost per } \Delta\text{KPI} \;=\; \frac{C_k - C_{k-1}}{K_k - K_{k-1}}$$

```
tb = input_state['tier_budget']
order = ['core', 'mainline', 'stretch']

rows, prev_cost, prev_kpi, prev_tier = [], 0.0, 0, None
for t in order:
```

```
    cost, kpi = tb[t]['cost'], tb[t]['kpi_delta']
    delta_kpi = kpi - prev_kpi
    marginal = (cost - prev_cost) / delta_kpi if delta_kpi > 0 else
    float('inf')
    rows.append({
        'tier': t,
        'cost': cost,
        'cum_kpi': kpi,
        'avg_cost_per_kpi': round(cost / kpi, 1) if kpi else None,
        'marginal_cost_per_kpi': round(marginal, 1),
        'marginal_from': prev_tier,
    })
    prev_cost, prev_kpi, prev_tier = cost, kpi, t

printable = all(rows[i]['marginal_cost_per_kpi'] <= rows[i +
    1]['marginal_cost_per_kpi']
                for i in range(len(rows) - 1))

output_state['tier_analysis'] = rows
output_state['marginal_cost_rising'] = printable
output_state['avg_vs_marginal_gap'] = round(
    rows[-1]['marginal_cost_per_kpi'] - rows[0]['avg_cost_per_kpi'],
    1)
```

Marginal cost can rise or fall as a program expands. Rising cost may reflect harder-to-reach participants or added capacity; falling cost may reflect shared setup costs, spare capacity, or an optimistic outcome assumption. Neither pattern alone demonstrates double-counting. In this invented example, marginal cost falls across the tiers, so the causal assumptions behind the extra transitions deserve scrutiny. Separately, S10 flags unclear credit allocation and double-counted impact as philanthropic funding problems. Check that the tier counts refer to distinct cumulative outcomes and that any attribution to multiple funders is handled explicitly. P07 reports career-role transitions and fellowship transitions separately; retaining that distinction makes the reporting clearer but does not, by itself, prove that the underlying groups cannot overlap.

In this example, the mainline tier costs $41,080 and assumes eight transitions, giving $5,135 per raw transition. That is an arithmetic result from assumed inputs. Whether it is worthwhile depends on causation, the value of the outcomes, omitted costs, and the relevant funding alternative.

### Step 3

The last move inverts the calculation into the question the funder is actually asking: how many outcomes must materialise for this grant to clear the bar. That requires a value per outcome and a bar multiple, and here the corpus is unusually instructive because the two candidate numbers are three orders of magnitude apart.

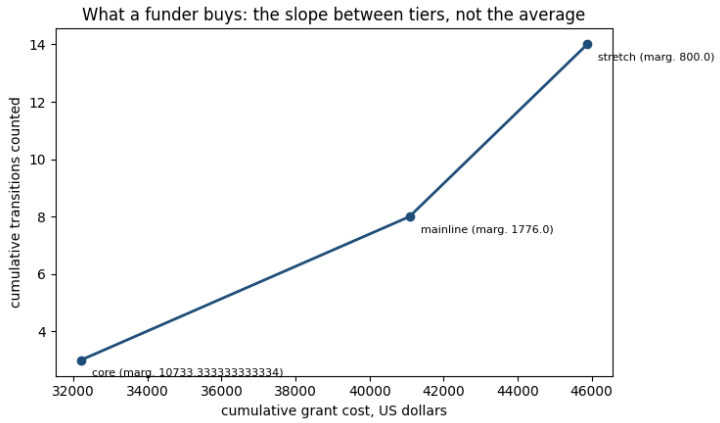


Figure 19: Graphical analysis for Step 2.

P05 reports an average weighted return on investment of roughly 5x to 6x for a portfolio of effective-giving organizations that moved about $300 million in 2025. P02 reports a mean modelled social return on investment of roughly 5,500x for a career-transitions portfolio. Those are not in conflict, because one is dollars moved to cost-effective charities and the other is modelled social benefit per dollar granted to a careers organization, but a memo that cites "the bar" without saying which quantity it means is not calibrated. **The break-even calculation is the honest way to handle this**: state the bar as a parameter, show the break-even across a range, and let the reviewer supply their own bar rather than smuggling one in.

In the supplied AI Governance RFP, D04 describes $200,000–$2,000,000 per year over one to two years as a usual starting range, with exceptions. It also welcomes applications below $200,000 per year when other funders have rejected the proposal. Our illustrative stretch tier is $45,880, so it falls below this preference; the preference does not make the ask categorically ineligible. D04 does not establish the same range for the Effective Careers program, or establish that this particular field-building project fits either program. P10 describes a different fund that usually makes grants below $150,000. These are possible funding routes to assess, not evidence of eligibility or a predicted award.

D03 describes BlueDot's typical rapid grants as up to $10,000. Each complete tier here exceeds that amount, including the $32,200 core. The captured guidance does not establish a hard universal cap or an exception for this example, so it supports neither categorical ineligibility nor an expectation that BlueDot would fund a whole tier. A smaller separately costed component might be considered, subject to the program's remit and decision; do not treat that possibility as funding already secured. D03 also says stipends and living expenses are not its usual scope, which matters for this labour-heavy example.

**Output status:** The cost and attribution tables above are the completed arithmetic. The funding break-even table is an unfilled exercise until V and r are specified. No funded year-one tier is recommended on the available information; mainline is the worked illustration, not an endorsed choice. In code-oriented use, return an explicit "undetermined—missing justified outcome value and threshold" status rather than pretending that a recommendation was calculated.

1. State the bar as a parameter with its provenance, and say which quantity it refers to.
2. Give break-even outcomes per tier at that bar, and at one bar on either side.
3. Name the funder whose threshold each tier is built to clear, using published thresholds where they exist.
4. State what evidence would justify the renewal at a larger ticket, in the terms D04 would assess it: theory of change, track record, strategic judgment, project risks, cost-effectiveness, scale.

A caution about the whole exercise: the tier ladder presupposes that the applicant can run the smaller tier well. L01's rationale for tiers is that a reviewer needs a menu of options quickly; it is not a mechanism for selling a project that cannot execute the base case. If the core tier cannot be delivered by the named accountable role, no amount of tiering fixes it.

## Sub-Question 6: Where the method is informal, uncertain or misleading, and how to disclose it

Stress the worked example, find which parameter actually drives the answer, and write the disclosure paragraph that separates what is source-supported from what is the applicant's inference or proposal.

### Step 1

A cost-per-outcome number carries spurious precision. The parameter uncertainty in the worked example is large and unevenly distributed, and the way to show that without a Monte Carlo simulation is a one-at-a-time swing chart: vary each parameter across its plausible range, hold the others at base, and rank by the size of the swing. The ranking is the finding. If the dominant parameter is the one nobody has measured, the model's decimals are decoration.

**Exercise:** Supply the named input explicitly before running this code. No result from this exercise has been calculated here.

The multi-parameter sensitivity calculation is an unfilled exercise. Supply coherent base inputs and parameter ranges before using it to rank sensitivities. The separate two-point chart illustrates only attribution sensitivity in the mainline example.

```
def model(params):
    labour = params['organiser_weeks'] * params['organiser_week_cost']
    events = params['events'] * params['cost_per_event']
    support = params['advisee_months'] *
    params['cost_per_advisee_month']
    cost = (labour + events + support) * params['overhead_multiplier']
    attributable_units = params['transitions'] *
    params['counterfactuality']
    return cost / attributable_units if attributable_units > 0 else
    float('inf')

base_params = input_state['sensitivity']['base']
ranges = input_state['sensitivity']['ranges']

base = model(base_params)
rows = []
for key, (low, high) in ranges.items():
    p_low, p_high = dict(base_params), dict(base_params)
    p_low[key], p_high[key] = low, high
    v_low, v_high = model(p_low), model(p_high)
    rows.append({'parameter': key,
                 'low': round(v_low, 1), 'high': round(v_high, 1),
                 'swing': round(abs(v_high - v_low), 1)})
rows.sort(key=lambda r: -r['swing'])

gross_cost = (base_params['organiser_weeks'] *
    base_params['organiser_week_cost']
              + base_params['events'] * base_params['cost_per_event']
              + base_params['advisee_months'] *
    base_params['cost_per_advisee_month']) \
             * base_params['overhead_multiplier']
value_per_unit = input_state['sensitivity']['value_per_unit']
target_multiple = input_state['sensitivity']['target_multiple']
output_state['tornado'] = rows
output_state['base_cost_per_unit'] = round(base, 1)
output_state['dominant_parameter'] = rows[0]['parameter'] if rows
    else None

# value_per_unit: value of one attributable transition, in the same
    value
# units as gross_cost. target_multiple is the required value/cost
    ratio.
# counterfactuality is the attributable share of the raw transition
    count.
counterfactuality = base_params['counterfactuality']
planned_raw_transitions = base_params['transitions']
if value_per_unit <= 0 or target_multiple <= 0:
    raise ValueError('Specify positive value_per_unit and
    target_multiple.')
```

```
if not 0 < counterfactuality <= 1 or planned_raw_transitions <= 0:
    raise ValueError('This example requires a positive attribution
    share <= 1 and a positive transition count.')

required_attributable_transitions = gross_cost * target_multiple /
    value_per_unit
required_raw_transitions = required_attributable_transitions /
    counterfactuality

output_state['break_even_units'] = round(required_raw_transitions, 2)
output_state['break_even_attributable_units'] =
    round(required_attributable_transitions, 2)
output_state['break_even_vs_plan'] = round(required_raw_transitions /
    planned_raw_transitions, 2)
```

Let B be the grant cost, T the projected raw transitions, c their attributable share, V the value per attributable transition, and r the required value-to-cost multiple. Then attributable transitions are T × c. The project reaches the chosen threshold when T × c × V ≥ B × r. Required raw transitions are therefore B × r / (V × c), and the comparison with the plan is B × r / (V × c × T). A ratio of 1 reaches this assumed threshold exactly; a larger ratio requires more raw transitions than planned. These statements depend on a defensible c, V, and r expressed in compatible units. None is established by the cost table alone.

| Tier | Assumed raw transitions | Attributable transitions at c = 0.3 | Cost per attributable transition at c = 0.3 | Attributable transitions at c = 0.8 | Cost per attributable transition at c = 0.8 |
|---|---|---|---|---|---|
| Core | 3 | 0.9 | $35,777.78 | 2.4 | $13,416.67 |
| Mainline | 8 | 2.4 | $17,116.67 | 6.4 | $6,418.75 |
| Stretch | 14 | 4.2 | $10,923.81 | 11.2 | $4,096.43 |

Fractional attributable counts here are expected values. The 0.3–0.8 attribution range comes from the original illustration, not measured project data. For any chosen funding multiple r, multiply a table entry by r to obtain the minimum value V per attributable transition needed to meet that threshold. A numeric verdict remains unfilled because the example does not establish V or r.

### Step 2

The disclosure paragraph is the memo's equivalent of the index the overall deliverable requires: a statement of which sections are source-supported, which are inference, and which are proposal. Producing it as a counted artifact rather than a sentiment is a small computational act with a real effect on the reader's trust calibration.

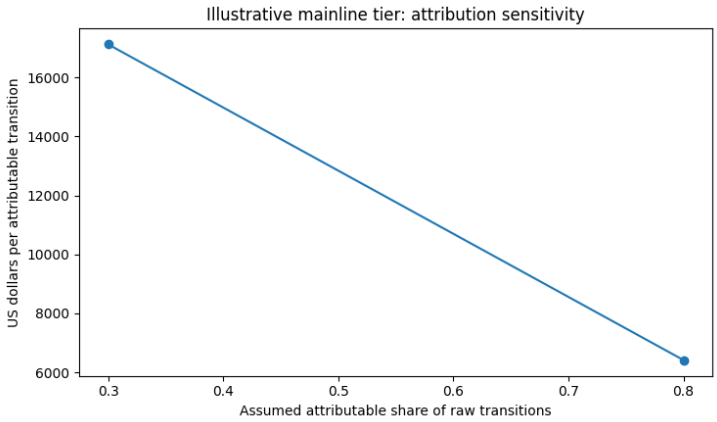


Figure 20: Graphical analysis for Step 1.

**Exercise:** Supply the named input explicitly before running this code. No result from this exercise has been calculated here.

```
tags = input_state['section_tags']

counts = {}
for section, tag in tags.items():
    counts[tag] = counts.get(tag, 0) + 1

total = sum(counts.values())
shares = {k: round(v / total, 3) for k, v in counts.items()}

disclosure = (
    'This memo has ' + str(counts.get('source_supported', 0)) + ' sections that rest on cited sources, '
    + str(counts.get('inference', 0)) + ' sections that are our inference from those sources, and '
    + str(counts.get('proposal', 0)) + ' sections that are proposed method rather than established practice. '
    + 'The budget uses the WBS/CBS/OBS/RACI decomposition as a proposal; we have no evidence that it is '
    + 'the template in use at any specific funder.'
)

output_state['tag_counts'] = counts
output_state['tag_shares'] = shares
output_state['disclosure_sentence'] = disclosure
```

- L01 provenance: a short personal note in Polish, self-described as an account of standard practice reconstructed from reading done long ago, with

the author unsure whether anything has been omitted. Treat the method as a proposal, not a house standard.

- L01's claim that this structure is uncommon in EA organizations is itself a claim about EA organizations, not evidence about how a given funder writes memos, and it is unfalsifiable from this corpus.
- L01's permission for a vibes-based KPI sits in direct tension with D01's requirement to state confidence and support. The resolution is to label the measure as an ordinal proxy, say who supplied the scale, and say what would turn it into a cardinal measure.
- The single-driver rule for cost pools is a simplification that fails for shared fixed costs. L01 flags the resource-based alternative as less informative, but resource-based costing is the common convention in practice, so a memo using cost pools is choosing legibility over familiarity and should say so.
- L02's worked examples contain explicitly invented premises and invented performance and publication claims. Borrowing the argument form is legitimate; borrowing the content is not, and the memo should mark invented illustration wherever it appears.
- The indirect evidence that a funder values this decomposition is D02's report that CG built memo-drafting and back-of-envelope-calculation tooling and treats calculations as a calibration instrument. That is an inference about institutional practice from a post about AI adoption, not a published rubric.
- The break-even bar is a parameter. Published anchor points differ by roughly three orders of magnitude across methods, so a memo that says "the bar" without specifying the quantity is uncalibrated even when the arithmetic is right.
- Intermediaries complicate the counting. If a small first funder seeded an organization that a large funder later scaled, the same outcome appears in two portfolios, and the correct treatment of credit is an open question rather than a settled method.

One further disclosure belongs in the memo rather than in this note. P04 supplies three distinct examples. Trevor Levin describes spending two hours finishing one write-up on a representative Tuesday. Chris Bakerlee reports investigating 6–12 grants at different levels of attention. Abbey Chaver describes dividing her time roughly equally among meetings, research, and investigations or memos. These observations come from different people and do not establish a two-hour end-to-end grant investigation or a universal writing budget. Levin's related advice is to seek decision-relevant evidence efficiently while avoiding unnecessary work for the organizations being evaluated. **A memo built on a four-structure decomposition can defeat its own purpose if it transfers the analytic work back to the reviewer**. The WBS, CBS, OBS and

RACI tables are instruments for the applicant's own clarity; what reaches the reviewer is a linked summary, a ranked list of the considerations that matter, the falsifier, the cost pools with their drivers, and the counterfactuality-adjusted cost per outcome.

# Map fund intermediaries, regranting and small-grant ecosystems.

## Question

Which vehicles sit between a donor's dollar and a working project, and how should an evaluator decide when to route money through a pooled fund, a regrantor, a matching facility, a recommender, or directly to an organization?

The unit of analysis is not the grantee organization but the **placement of decision rights**: who holds taste, who holds money, who holds dealflow, and who carries the residual when the grant fails. The deliverable is a map plus routing rules, not a portfolio recommendation.

## Background

The corpus supplies a natural decomposition. P12 treats any grant program as a bundle of five inputs: **Money, Taste, Dealflow, Hustle, and Trust**. A vehicle is then defined not by its name but by which inputs it supplies and which it delegates. This is the load-bearing analytic move: it converts a list of organizations into a matrix of delegations.

P12 also names the underlying failure mode as the ***cursed game of philanthropy***: the principal (Money) cannot cheaply verify the agent (Taste), so every vehicle in this map is best read as a different answer to the same principal-agent problem. Treat trust as the scarce input, with money second.

*Scope caveat. The source posts are dated 2025-2026. LTFF has wound down, the Transformative AI Fund has just launched with a seed of roughly $1M, and BlueDot and ACX round sizes are reported for past cycles. Everything below is a descriptive reconstruction of the corpus and is not current-cycle guidance.*

## Sub-Question 1: The topology: what is an intermediary of what

The intuitive map is a line: donor, then fund, then grantee. The corpus supports a two-axis map instead. Axis one is **where decision rights sit**; axis two is **which stage of the funding funnel the vehicle serves**. A vehicle can be high on one axis and low on the other, and the interesting cases are the off-diagonal ones.

Three roles are routinely conflated and should be kept distinct: grantmaker, recommender, and regranter. Coefficient Giving occupies all three simultaneously, which is the single most confusing feature of the landscape.

**Step 1**

Define each role by its delegation profile rather than by its branding. A role is identified by the inputs it supplies, which inputs it outsources, and where the loss lands when the grant fails.

- Direct giving. The donor holds all five inputs. No delegation, no verification cost, and near-full overhead on every grant (P10 notes that a solo donor pays close to full overhead per grant).
- Pooled fund. Donors delegate Taste, Dealflow, and Hustle, and bank Trust in the fund itself. P10's stated advantages are exactly these three: dealflow, expertise, operations and tooling. P10 explicitly carves out the exception, namely well-informed and well-connected donors who may beat fund recommendations.
- Regrantor. Money is delegated to part-time agents who already hold Taste and Dealflow; a platform supplies Hustle and a public transparency layer (P11, P12). This is delegation of the pot, not of the judgment.
- Matching facility. The upstream funder holds Taste provisionally and delegates the marginal dollar to a crowd, using a threshold as a commitment device (N02).
- Recommender. The recommender holds Taste and Dealflow but not Money. D04 is explicit that a strong application might be funded by Good Ventures or by more than 20 other philanthropists, that these philanthropists give far less each year than CG, that they are typically interested in only a narrow subset of opportunities, and that in aggregate they do not act on most recommendations.
- Regranter. An upstream funder delegates to a downstream evaluator with better local taste, for example CG to GiveWell. P06 defends this as impactful but says CG does it only when it sees a distinctive case.

The non-trivial implication: when a grantmaker recommends a vehicle, they are simultaneously deciding **who holds the residual**. Recommending a fund moves the residual to the fund; recommending a regrantor moves it to a named individual; recommending direct giving keeps it with the donor. Most disagreement about vehicles is disagreement about who should carry that residual, which is not the same as disagreement about expected value.

**Step 2**

Assemble the map as a single reference table. Sizes are the source claims, rounded, and many are retrospective or projected.

Table 40: *Descriptive map assembled from P06, P10, P11, P13, P14, P15, P16, P17, D03, D04 and N02. Sizes are source claims from 2025-2026 and should not be read as current terms.*

| **Vehicle** | **Decision-rights locus** | **Typical ticket** | **Stage it serves** | **Transparency** | **Key caveat** |
|---|---|---|---|---|---|
| CG thematic funds | CG holds Taste and Money; partners add Money | first grants usually 200k to 2M/yr; over 20M for high conviction | renewals and large proactive seeds | public grants database; D01 reasoning-transparency norms | D04 says projects below 200k/yr should ideally be funded elsewhere |
| CG as regranter | downstream evaluator holds localized Taste | large blocks | mature, evidence-backed portfolios | downstream public writeups | P06: justified only where there is a distinctive case |
| CG as recommender | CG holds Taste; external philanthropist holds Money | varies; sharing permission is the gate | any opportunity CG likes but will not fund itself | applicant controls permission to share materials | D04: more than 20 philanthropists; most recommendations are not acted on |
| EA Funds Transformative AI Fund | full-time team holds Taste; donors hold Money | typically below 150k; rarely above 300k; nothing near 1M for now | individuals, new orgs, existing orgs with new projects | transparent by default; quarterly reports; private grants aggregated only | seed of roughly 1M at launch; tells donors to judge it by later quarterly reports |

| Vehicle | Decision-rights locus | Typical ticket | Stage it serves | Transparency | Key caveat |
|---|---|---|---|---|---|
| LTFF (legacy) | person-bound Taste; wound down with its team | 820+ grants, just under 35M since 2017, roughly 42k average | first checks for people and projects | grants database | a fund whose identity is its team is not an institution; it terminates with the team |
| Manifund regranting | Money colocates with a part-time regrantor's Taste and Dealflow | 2025: 13 regrantors sharing 2M, about 150k per pot | early, contrarian, pre-institutional | full public dashboard of every regrant | the product is regrantor selection, so the program is only as good as its roster |
| BlueDot Rapid Grants | community-scoped Taste; fast operational Hustle | 50 to 10k, most a few hundred; about five working days | compute, events, community, career acceleration | public grantee list; light impact update instead of a report | eligibility is community-linked, so the grant doubles as a retention instrument |
| ACX Grants | one person's Taste plus volunteer evaluators; audience supplies Dealflow | round of about 1M; most grants 5k to 50k; rare to 100k | novel, weird, systems-changing projects outside standard categories | public results post with names and amounts | SAFEs and convertible grants keep an upside claim; the seal of approval is rival |

| Vehicle | Decision-rights locus | Typical ticket | Stage it serves | Transparency | Key caveat |
|---|---|---|---|---|---|
| SFF matched grants | upstream funder holds provisional Taste; crowd supplies the rest of the Money | N02: 26k goal, 1:1 match, 10k minimum to unlock | small teams with one funder already convinced | letter of intent published alongside the campaign | the match is a coordination device, not an independent evaluation |

**Step 3**

Resolve the double role of Coefficient Giving, because it is the part of the map most likely to be mishandled in an interview answer.

CG is upstream of other money (it recommends to more than 20 philanthropists, so it acts as a taste-and-dealflow intermediary for capital it does not control), downstream of other money (it accepts external donors and grew non-Good Ventures funding past 200M), and internally a regranter (GiveWell, the Regranting Challenge). A single institution therefore appears at three different positions on the map. The practical consequence is that **sharing permission** is a design variable inside every recommendation, not an administrative afterthought. D04 says applications are substantially more likely to be funded if applicants give permission to share, which means a grantmaker's recommendation has an option value that depends on a clause the applicant controls.

Calibrated formulation to reuse: *CG is a mega-grantmaker, a small-grant router, and a recommender at once, so the question is never whether to use CG or an intermediary but which of CG's three functions a given opportunity should pass through.*

## Sub-Question 2: Scale bands, the small-check gateway, and why the market has seams

Do the vehicles form a smooth continuum of ticket sizes from a few hundred dollars to hundreds of millions, or a set of discontinuous bands with seams between them? The answer determines whether a project can present one budget or must be reshaped to fit a funder.

**Step 1**

Assemble the numeric bands before drawing any conclusion, because the shape of the distribution, not the identity of the funders, carries the argument.

1. BlueDot Rapid Grants: 50 to 10k, most grants only a few hundred dollars for compute and API credits, decisions in about five working days (P13, P14, D03).
2. Manifund regranting: individual regrants observed in the top-hits list span about 2.5k to 335k, with a 2025 pool of 2M across 13 regrantors, roughly 150k per pot (P11).
3. ACX Grants: about a 1M round, most grants 5k to 50k, a rare few up to 100k, 10 to 50 grantees (P16).
4. Transformative AI Fund: grants typically below 150k, above 300k rare, and nothing near 1M for the foreseeable future (P10).
5. LTFF legacy scale: 820+ grants totaling just under 35M since 2017, implying a mean near 42k (P10).
6. CG: D04 says first-time grants are typically 200k to 2M per year over one to two years, and smaller projects should ideally be funded by other funders such as SFF, LTFF, Founders Pledge, the AI Risk Mitigation Fund, and Longview.

**Step 2**

Reconstruct the gateway argument in its strongest form and then in its weakened form. P10's version is that large checks to established organizations presuppose the initial small checks that got those organizations started, that organizations take time to build, and that the next 12 to 18 months disproportionately determine whether the ecosystem reaches a useful size in time.

The non-trivial point is that this is ***not a time-value-of-money claim***. At a zero discount rate the argument still holds, because the object being bought is absorptive capacity at the moment the funding wave arrives. Formalize it that way and the argument becomes falsifiable rather than rhetorical.

$$A(T) = F\big(t_{\text{funding}}\big), \qquad \frac{\partial A(T)}{\partial t_{\text{funding}}} > 0 \text{ at } r = 0$$

where $A(T)$ is the field's capacity to absorb capital at the arrival date of the funding wave. The weakening move, which P10 concedes, is that the derivative is only positive for organizations that would otherwise choose a safer option; for organizations that would be founded anyway, the early check is a pure transfer.

**Step 3**

Explain the seams. Ticket bands are set by the funder's operating cost per grant, not by the project's natural size. A fund that replies within days cannot spend hours on a 20k request, and a funder that runs full grant investigations cannot move at 3k. The result is a lumpy market with a visible handoff region

between roughly 150k and 200k, exactly where the small funds stop and CG's stated floor begins.

Four reshaping moves follow, and recognizing them is the practical payoff of the map: **sequence** (take a rapid grant now, return for the larger tranche with traction), **pad** (expand scope to reach a band's floor without inventing work), **split** (separate the compute budget from the salary budget so each can clear a different vehicle), and **bundle** (coalition of several small grantees or a fiscal sponsor). An evaluator who cannot name these moves will misread a good project as unfundable.

*Bounded caveat: the seams argument is an inference from operating-cost logic, not a claim any source makes directly. Present it as a hypothesis with the band data as support.*

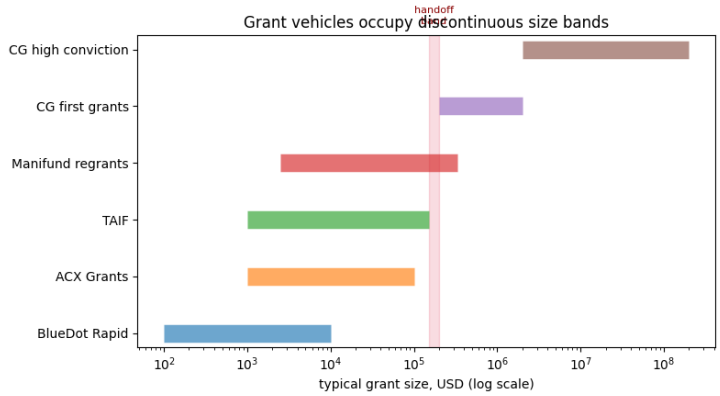


Figure 21: Graphical analysis for Step 3.

## Sub-Question 3: Trust, taste, dealflow, hustle and money: the central disagreement

Is dispersing small pots to many part-time grantmakers better than centralizing taste in a professional grantmaking organization? The corpus contains both positions, held by people with no reason to be insincere, and the disagreement should be preserved rather than resolved into a single voice.

### Step 1

Fix the five terms precisely, because loose usage is what makes the disagreement look like a personality clash.

- Money: the budget to give away, historically from earning-to-give individuals, then Good Ventures, the FTX Future Fund episode, and now expected Anthropic and OpenAI wealth (P12).

- Taste: the capacity to distinguish good from bad projects, subdivided into taste in people, in ideas, and in fields across cause prioritization (P12). Its awkward property is that everyone believes they have it and there is no single true taste.
- Dealflow: the opportunities that reach you, distinguished into seeing projects others would not, seeing more of them, being preferred by good opportunities, and creating opportunities from scratch (P12).
- Hustle: the coordination glue between money, taste and dealflow, including logistics, follow-ups, and payment rails (P12).
- Trust: the meta-input. P12 predicts trust will be the bottleneck to deploying the coming wave of funding, having worked so far only because the community was small, weird, demanding and socially interwoven, a condition that does not scale. Ben Kuhn's point that trust bottlenecks team growth is quoted as the analogy.

**Step 2**

Notice that the corpus offers **three competing institutional answers to the same trust problem**, and this is the analytically interesting object. They differ in where they place the unverified party, and each fails differently.

1. Colocation. Put Money in the same person as Taste and Dealflow. Manifund regranting does this deliberately (P11, P12), and Leo Gao's microgrants from personal funds are the pure case. Failure mode: only as good as the roster, and each pot is capped by one person's bandwidth.
2. Decision-with-money. Strip the agent of final authority, so the principal makes the last call on the evaluator's recommendation. Lightcone Commons is the cited example of a design that requires less trust by placing final funding decisions with money (P12). Failure mode: it verifies taste by demanding legibility to the principal, which is exactly what taste cannot supply.
3. Hiring. If you have moderate trust in someone's taste, buy their Taste, Dealflow and Hustle outright (P12's closing advice). Failure mode: cost, and the talent market. P06 notes CG competes with AI labs that can pay millions, and is considering pay changes in response.

The non-obvious comparison is between the first two. Colocation and decision-with-money are not two points on a spectrum of trust; they are opposite directions of travel. Colocation accepts that taste is unverifiable and embeds it in the decision-maker. Decision-with-money accepts that money is unverifiable-by-proxy and embeds verification in the principal. A grantmaker who invokes trust as a bottleneck should be able to say which direction they are moving.

**Step 3**

State the disagreement in its own terms, then locate the crux.

- For dispersion (P11). Many people's first 100k of grants will beat a professional grantmaker's last 100k. Any grantmaker's field of view is bounded by network, geography and expertise, so dispersed part-timers notice things a central team cannot. It is far more scalable, because giving someone a regranting budget is lower friction than hiring, training and retaining a professional. It builds a talent pipeline: Joel Becker moved from regrantor to METR staff.
- For centralization (P12 and P06). P12 describes CG as a small army of specialized grantmakers, an army of operations people, and trust that began inherited from GiveWell and now stands as an institution in itself. P06 defends the staffing ratio explicitly: CG moves more money and makes more grants per staff member than comparable philanthropies, and regranting is used only where a distinctive case exists, in part because moving grantmaking staff onto a grantee's payroll is not obviously an efficiency win.
- Where the disagreement is empirical rather than about values. The two claims are about different positions on a returns curve. Dispersion wins where the marginal opportunity is outside the central team's field of view, that is, at the frontier of a growing field. Centralization wins where the marginal opportunity is inside a mapped field and the binding constraint is consistency, documentation, and portfolio-level reasoning. P10's own framing supports this reading: transparency, reliable responses and open applications matter most for discovering unsolicited opportunities.
- Where it is genuinely about values. Taste pluralism is not resolvable by evidence. P12 says outright that there is not necessarily one true taste, and that money might instead consider whether a particular taste aligns with her values. S11 is the same point from the other side, listing an explicit set of disagreements with an AI safety consensus on open weights, secrecy, China, and philanthropic funding. A fund that concentrates taste thereby selects a worldview.

Synthesis to use in an interview: the dispersed-versus-central argument is underspecified until you say which margin you are funding. Offer the conditional: *disperse when searching, centralize when consolidating, and be explicit that concentrating taste is a worldview choice rather than a purely technical one.*

## Sub-Question 4: Routing rules: when to recommend each vehicle

Given a stated donor profile, which vehicle should be recommended, and what would change the recommendation? Produce rules that are derived from the

delegation logic rather than from brand preference.

**Step 1**

Elicit six parameters before recommending anything. A recommendation that does not name these is not yet a recommendation.

1. Check size per year and the number of grants the donor is willing to make. This sets overhead per dollar and therefore the fund's cost advantage.
2. The donor's own dealflow, measured as whether good opportunities currently reach them unsolicited.
3. The donor's own taste in people and fields, and their confidence in it. P10 concedes that well-informed, well-connected donors may beat fund recommendations.
4. Time available for hustle, since the fund absorbs almost all of it.
5. Failure tolerance and whether the donor accepts power-law payoffs in which most grants return little and one return pays for the portfolio (P10).
6. Transparency preference and appetite for shaping versus responding, since P06 names the strategic funder trap and P05 names how much to shape the ecosystem as an open question.

**Step 2**

Apply the following routing table. Each row names the binding constraint and the recommendation that follows from it, together with what would flip the answer.

Table 41: *Routing table derived from the delegation logic in P10, P11, P12, P15, P16 and N02. Profiles are illustrative; the reasoning transfers, the specific names may not.*

| **Profile** | **Binding constraint** | **Recommend** | **Why** | **Flip condition** |
|---|---|---|---|---|
| Earning-to-give at 50k/yr, little domain knowledge | No dealflow, no verified taste, high hustle cost | Pooled fund (TAIF or an EA Fund) | P11 names this profile directly: a solo donor without dealflow funds what is easiest to find, which is disproportionately what least needs a marginal dollar | Donor can name three opportunities nobody told them about |
| Lab employee or researcher with taste but little liquidity | High taste and dealflow, low money | Regrantor, or a personal microgrant program | Colocation puts money exactly where the unverifiable taste already sits (P11, P12) | No demonstrated track record or no dealflow beyond their own lab |
| Donor with a strong specific prior on one mature organization | The binding constraint is not discovery but conviction | Direct giving, or general operating support | A fund's dealflow advantage is redundant here, so its overhead is a deadweight; D04 notes that general support criteria are stricter precisely because the assessment is broader | The prior is inherited rather than examined |

| Profile | Binding constraint | Recommend | Why | Flip condition |
|---|---|---|---|---|
| Donor who wants optionality and a recycled corpus | Wants upside if a grantee becomes a company | ACX-style round with SAFEs and convertible grants | P16 argues some prior grantees became valuable companies, and leaving that money on the table is wasteful; the terms are designed to err in the grantee's favor | Legal and conflicts burden exceeds the corpus value; P17 flags conflicts of interest once a portfolio exists |
| Donor who wants to push a specific small team over a threshold | A conditional match needs a last dollar | SFF-matched donation | N02: 26k goal matched 1:1 with a 10k floor, so a small gift unlocks a large block; the marginal donor captures the entire discontinuity | The match would be reached anyway, or the upstream funder's diligence is not independently credible |

**Step 3**

Worked exercise, stated as proposed practice rather than as a source finding: compare direct giving, giving to a fund, and giving to a regrantor for a donor with 500k per year, ten hours per week, moderate domain knowledge, and high failure tolerance. Model expected value as the probability of a top-decile hit times the value of a hit times the retained fraction after overhead.

$$\mathbb{E}[V_{\text{vehicle}}] = p_{\text{hit}} \cdot V_{\text{hit}} \cdot (1 - c_{\text{overhead}}) \cdot B$$

The disputed parameter is $p_{\text{hit}}$, and the honest treatment is to report the crossover rather than a point estimate. The code below computes the crossover in dealflow quality at which direct giving overtakes each delegated vehicle, which

is the only form of the comparison that survives disagreement about the parameters.

## Sub-Question 5: Matching, credit allocation, and why impact does not add up

How do matched grants and multi-stage funding chains distort measured impact, and what convention should an evaluator use so that reporting stays honest?

### Step 1

Model the match as a discontinuous payoff rather than as leverage. In N02 the campaign needs 26k, the match is 1:1, and at least 10k must be raised to unlock it. The value of a donor's dollar therefore depends on where the total lands, not on the donor's own amount.

$$V_{\text{donor}}(R) = 1 \; + \; C \mathbb{1}[\, R_{\text{final}} \geq M \,]$$

Here $R_{\text{final}}$ is total funds raised, $M$ is the unlock threshold, and $C$ is the unlocked block. The donor who supplies the last marginal dollar before the threshold captures the whole discontinuity. That is a strong incentive to be last, which is a coordination failure, and it is exactly the pattern S10 names as a game of chicken between funders that leads to glacial response times. Symmetrically, an early donor who is not certain the threshold will be reached is buying an option, not a grant.

*Practical read for an evaluator: a matched grant is not an independent quality signal. The upstream funder's letter of intent conveys their diligence, which is useful, but the match mechanic itself conveys only that a threshold was crossed.*

### Step 2

Examine the credit problem concretely. P11 reports that a 143k Manifund regrant seeded Timaeus, which later merged with Resolution and received a 160M CG grant, that the implied stake grew 53x to about 7.6M, and that this single grant would pay back the entire regranting budget to date of 5.2M. P12 gives further examples, including Apollo Research raising a seed round said to imply roughly a 10x impact return on a 335k regrant.

Three things are true at once here, and conflating them is the standard error. First, the counterfactual is doing the work: the claim is that Timaeus would not have started without the seed, which the source asserts but does not demonstrate. Second, the returns are non-additive across vehicles. If both the regrantor and CG claim the Resolution outcome, the sum of claimed impact exceeds total impact. Third, the ROI figures are option values at a point in time, not realized returns, so they can move by an order of magnitude.

$$\sum_i \widehat{\text{impact}}_i \gg \text{impact}_{\text{total}} \quad \text{whenever} \quad \widehat{\text{impact}}_i = \text{impact}_{\text{counterfactual given } i}$$

The same structure explains why portfolio returns are the only meaningful aggregate. Under a power law, $\max_i V_i \gg \sum_{j \neq i} V_j$, so an ex-post hit cannot validate the ex-ante judgment that selected it, and a run of misses cannot refute a good selection rule. P10 states the same point: expect power-law-distributed payoffs, most grants return little or nothing, and what pays for the portfolio is the occasional home run.

### Step 3

Adopt an attribution convention for written work. It should be short enough to apply under time pressure and honest enough to survive a challenge.

1. State the counterfactual explicitly and mark whether it is supported or assumed. Directly state the counterfactual, mark whether it is supported or assumed. If the claim is that an organization would not exist without this check, say so and give the basis.
2. Separate origination from scaling. Origination credit belongs to whoever made the entity exist; scaling credit belongs to whoever supplied the later capital. Do not report a combined number.
3. Report ranges, and label option values as option values. Any multiple derived from a subsequent round is a mark, not a return.
4. Never sum across vehicles. Present the chain as a chain, with who bore what risk at each link, which is also how P12 frames the trust question behind principal and agent.
5. Note the residual holder. If the fund carries the reputational cost of a miss and the regrantor carries the origination risk, the write-up should say which one the evaluator is judging.

## Sub-Question 6: Failure modes, caveats, and what would change this map

Where does this map mislead, what is inference rather than source, and which observations would force a revision? Finish with an interview-ready formulation and a rehearsal drill.

### Step 1

Name the failure modes of the intermediaries themselves, not of the map.

- Band drift. A small fund succeeds by staying in its niche and then grows into a different vehicle. P10 commits the Transformative AI Fund against this by saying it will not make grants above roughly 1M for the foreseeable future because its comparative advantage is not with established organizations.
- Person-bound institutions. LTFF wound down because its identity was tightly bound to the individuals running it. A fund whose trust is not institutional is not a durable vehicle, which sharpens the contrast with P12's account of CG's trust standing as an institution.
- Capacity collapse under volume. D04 records that CG raised the bar for which EOIs and full proposals to invite forward because applications overwhelmed staff, so proposals previously invited would now be rejected. A volume-driven intermediary can silently reprice its own bar.
- The rival seal. P17 notes that nonfinancial prestige has zero direct cost, so a naive cost-benefit analysis favors granting it to everyone, which destroys its signal value, and the resulting paradox is that a 1 dollar grant may be harder to obtain than a 5,000 dollar grant. Any intermediary that trades in approval faces this.
- Transparency versus confidentiality. P10 says it will publish normally, but when it concludes with a grantee that work should not be publicized it will omit them from the database while including them in aggregate reports. P14 warns applicants to assume only written, project-specific confirmations count. The transparency default is therefore weaker than it appears at the level of individual grants.
- *Regrantor selection as a hidden single point of failure. P11's case rests on a roster, and the program is only as good as that roster; a donor choosing regranting is choosing people, not a process.*

**Step 2**

Separate source-supported content from inference, because this step contains a great deal of the latter and the distinction is itself part of the assessment criteria in D04, which lists reasoning transparency and scope sensitivity among the things evaluated.

- Source-supported: the five-component decomposition; LTFF's scale and wind-down; the Transformative AI Fund's stated scope and limits; Manifund's pool size and stated arguments; BlueDot's and ACX's stated terms; SFF's match terms in N02; CG's own statements about regranting and staffing ratio.
- Inference: that ticket bands are set by funder operating cost; that the 150k to 200k region is a structural seam; that the dispersed-versus-central

disagreement is conditional on position on a returns curve; that match mechanics create a last-dollar problem in this specific ecosystem.

- *Speculation, and should be flagged as such: that any named vehicle will continue operating at its stated scale, or that CG's stated floor will persist.*

Falsifiers worth naming. The seam hypothesis fails if a vehicle announces funding in the 150k to 200k region without reshaping the project. The gateway argument fails if a cohort of organizations is shown to have grown equally well from large first checks. The dispersion argument fails if regrantor outcomes are systematically worse than central outcomes on the same opportunity set. The colocation argument fails if regrantors with high taste but no platform support perform no better than random selection.

**Step 3**

Close with a rehearsal. The drill is to give a ninety-second answer to the question of when to use a fund versus direct giving, then check it against the four requirements implied by the role brief: reasoning transparency, quantitative reasoning, decisive considerations, and proportionate due diligence.

Model answer, calibrated and short: *A fund buys dealflow, expertise and operations, so it adds most value where a donor lacks the opportunities and least value where the donor already has conviction about a specific organization. Regranting buys the same things but places them in a named person, which solves the principal-agent problem by colocation rather than by verification, and carries roster risk instead. My decisive considerations are the donor's own dealflow and the size of the check relative to the fund's band; the failure mode I would watch is a fund drifting out of its niche as it grows. My confidence is moderate: the decomposition is well supported, but the claim that a vehicle is worth its overhead in a specific case is a judgment I would want to test against one or two comparable grants.*

# Practice judgment under uncertainty and generate alternative defensible answers.

## Question

Produce three judgment drills grounded in the supplied corpus. Each drill must present a realistic grantmaking or evaluation decision, at least two defensible answers, the assumptions each answer depends on, the decisive considerations, the crux, and what would change the answer. For each drill, also state what evidence is missing and label source-supported claims, inferences, and invented premises.

The drills should exercise the craft methods in this phase: reasoning transparency, probabilistic expected value, cost-effectiveness and social return modeling, hits-based portfolio logic, proportionate due diligence, theory-of-change assessment, grant-program design, and fund or regranting intermediation.

## Background

The role brief says evaluation requires comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and judging people and organizations. That is the target skill for these drills.

D01 supplies the reasoning-transparency rubric: open with a linked summary of key takeaways, indicate which considerations are most important, and indicate confidence in major claims and the kind of support for them. It warns that words such as plausible are read differently, so numeric or interval expressions are preferable where stakes justify the cost.

L02 supplies a six-move argument structure: meet the reader where they are; take the reader to where you are; show how you could be moved; show the evidence; show whether the evidence moves you; show where the reader should now be. Its practice notes identify which move carries too much weight and where evidence is still needed.

N14 supplies the political-feasibility warning: technically sound proposals can fail without coalitions, policy windows, and prior expert consensus. N06 supplies the partial-map warning: maps of governance activity are jurisdiction-dependent and incomplete.

Constraint: Do not invent personal credentials, work history, or an application. Label hypothetical or invented premises. Distinguish source-supported conclusions from inference. The drills below use invented scenario details only where they are clearly framed as hypothetical.

## Sub-Question 1: Drill 1: Big initial bet or staged pilot for a high-conviction new organization?

You are advising Coefficient Giving on a new AI safety research organization. The founder is a world-class technical researcher with no prior experience leading an organization. The proposal asks for $40 million over three years to hire researchers, buy compute, and build an institution.

Project Tailwind's published tiers include pre-seed of $200,000 to $2 million, seed of $2 million to $20 million, and scale of $20 million to $200 million or more. D04 says CG typically prefers to begin with $200,000 to $2 million per year over one to two years, then renew after evidence of success. P01 describes a $160 million grant to Resolution as CG's largest ever AI safety grant and says CG wants to bet big from the start where conviction is high. Decision: fund the $40 million ask now, or stage it?

**Step 1**

**Step 1: Generate two defensible answers and expose their assumptions.**

Table 42: Drill 1 comparison of two defensible answers.

| Element | Answer A: Fund at or near ask now | Answer B: Stage a pilot first |
|---|---|---|
| **Recommendation** | Fund $30-40 million over three years, with milestone-based tranches. | Fund $2-5 million for 12-18 months, then renew if evidence clears the bar. |
| **Core assumption** | The founder's technical judgment and recruiting ability are the bottleneck; organizational execution can be bought or learned; the field is capacity-constrained and high-conviction bets are inframarginal. | Organizational execution is a separate, testable capability; a smaller grant buys decision-relevant evidence; CG's staff time and reputational capital are better used to stage risk. |

| Element | Answer A: Fund at or near ask now | Answer B: Stage a pilot first |
|---|---|---|
| **Decisive consideration** | Whether the opportunity is inframarginal and time-sensitive, and whether delaying risks losing the founder or the project to another funder or employer. | Whether a pilot can produce evidence on hiring, research output, and management before a large commitment, at acceptable cost. |
| **Source support** | P01: majority of GCR grants over $20 million ever made were made in the current year; Resolution $160 million; P06: inframarginal grants often required recruiting a founder or incubating an organization. | D04: typical first grants $200,000-$2 million per year over one to two years; P04: Bakerlee tests assumptions and track record; P06: new grantees are labor-intensive per dollar early. |
| **Inference vs source** | Recommendation levels and tranche structure are inferences; source supports the existence of large high-conviction bets and the logic of inframarginal grants. | The specific pilot size and duration are inferences; source supports CG's stated preference for staged initial grants and evidence-based renewal. |

Both answers can be defended without contradicting the sources. The choice depends on which risk is larger: the risk of losing an exceptional founder by underfunding, or the risk of a large misallocation when organizational execution is unproven.

**Step 2**

**Step 2: Crux, reversal conditions, and missing evidence.**

- **Crux:** Is the binding constraint founder quality or organizational execution? If founder quality is the binding constraint and the founder is genuinely scarce, fund big early. If execution is the binding constraint and can be tested cheaply, stage.
- What would change the answer: evidence that the founder has successfully recruited senior researchers, managed a budget, or led a team; evidence

that another funder is about to make a competing offer; evidence that the organization's research agenda requires large compute or long time horizons that a pilot cannot support.

- Missing evidence: base rates of first-time founders in AI safety building durable organizations; the cost of losing the founder to a frontier lab or another funder; the marginal return on a dollar at the proposed scale versus a pilot; the specific milestones that would justify renewal.
- Six-move sketch: meet the reader with CG's stated preference for staged grants; take the reader to the inframarginal logic; show what would move you (pilot evidence versus founder-loss risk); show the evidence (analogies such as Resolution and MATS, but note differences); assess whether the evidence moves you; update to a recommendation that names the decisive condition.

Table 43: Drill 1 cruxes and reversal conditions.

| **Crux** | **Reversal condition** | **Label** |
|---|---|---|
| Founder quality vs organizational execution | If founder has strong evidence of execution, move toward Answer A; if not, move toward Answer B. | Inference from D04, P01, P06 |
| Time sensitivity and competing offers | If credible competing offer exists, increase front-loading. | Hypothetical premise unless verified |
| Scope of research agenda | If agenda requires large compute or institution upfront, move toward Answer A. | Inference from P01 and S09 tier structure |

## Sub-Question 2: Drill 2: Fund a regranting intermediary or build direct grantmaking?

You have $5 million to deploy into early-stage AI safety and governance. Option A: grant the $5 million to a Manifund-style regranting program that gives small pots to dispersed part-time grantmakers. Option B: use the $5 million to hire two or three full-time grantmakers and do direct grantmaking.

P11 argues regranting is fast, low-friction, and proactive, and reports that a $143,000 regrant helped start Timaeus, which later merged with Resolution and received a $160 million CG grant. P12 decomposes grant programs into money, taste, dealflow, hustle, and trust, and says trust is a likely bottleneck. P10 describes the Transformative AI Fund as primarily early-stage, typically below $150,000, and argues funds can have dealflow and operational advantages. S10 warns that philanthropic funding can suffer from funding based on cost

rather than value delivered, game-of-chicken response times, and unclear credit allocation. Decision: intermediary or direct?

**Step 1**

**Step 1: Generate two defensible answers and expose assumptions.**

Table 44: Drill 2 comparison of two defensible answers.

| Element | Answer A: Fund the intermediary | Answer B: Build direct capacity |
|---|---|---|
| **Recommendation** | Grant $5 million to a regranting program with transparent reporting and a portfolio approach. | Hire two to three grantmakers and run a direct early-stage program. |
| **Core assumption** | The binding constraint is dispersed taste and dealflow, not institutional capacity; many first $100,000 grants can beat a professional grantmaker's last $100,000; trust can be delegated to known regrantors. | The binding constraint is institutional control, counterfactuality, and credit allocation; direct staff can build durable field knowledge and avoid principal-agent problems. |
| **Decisive consideration** | Whether the intermediary has a track record of hits-based returns and whether its incentives align with CG's bar. | Whether CG can hire grantmakers with sufficient domain knowledge and whether the cost per successful grant is lower than through an intermediary. |
| **Source support** | P11: early grants can have huge returns; regranting is scalable; P12: money, taste, dealflow, hustle, trust; P10: funds can have dealflow, expertise, and operational advantages. | P12: principal-agent problems and trust bottleneck; S10: unclear credit and cost-based funding pitfalls; P04 and D04: direct grantmakers investigate, write up, and own decisions. |

| Element | Answer A: Fund the intermediary | Answer B: Build direct capacity |
|---|---|---|
| **Inference vs source** | The specific $5 million allocation and reporting requirements are inferences; source supports the mechanism and its rationale. | The specific hiring plan and program design are inferences; source supports CG's direct grantmaking model and its due-diligence criteria. |

Both are defensible. The intermediary answer buys breadth and speed; the direct answer buys control and institutional learning. The right choice depends on whether the bottleneck is grantmaker time or opportunity discovery, and on how much credit allocation and counterfactuality matter for the decision.

**Step 2**

**Step 2: Crux, reversal conditions, and missing evidence.**

- **Crux:** Is the binding constraint trust and dispersed taste, or is it CG's ability to process and originate grants? If trust is the bottleneck, intermediaries with aligned regrantors are attractive. If CG's own judgment is the scarce input, direct hiring is attractive.
- What would change the answer: evidence on the hit rate and counterfactual impact of past regrantors; a credible estimate of the cost per successful grant for each model; evidence that a regranting program would fund things CG would not want to be associated with; evidence that direct grantmakers cannot be hired at the required quality.
- Missing evidence: base rates of regrantor portfolios; attribution rules for later-stage CG grants to earlier regrantors; the administrative cost of the intermediary; the degree to which regrantors' taste aligns with CG's theory of change.
- Six-move sketch: meet the reader with the shared goal of deploying $5 million well; take the reader to the two models; show what would move you (hit-rate evidence, cost per grant); show the evidence (Timaeus example, but one example is not a portfolio); assess whether the evidence moves you; update to a recommendation that may combine a direct pilot with an intermediary allocation.

Table 45: Drill 2 cruxes and reversal conditions.

| Crux | Reversal condition | Label |
|---|---|---|
| Trust vs institutional control | If intermediary has strong aligned regrantors and transparent reporting, move toward Answer A; if principal-agent risk is high, move toward Answer B. | Inference from P12 and P11 |
| Dealflow vs processing capacity | If CG cannot hire, move toward Answer A; if CG can hire and retain, move toward Answer B. | Inference from P04, P06, P10 |
| Credit allocation and counterfactuality | If later CG grants would otherwise obscure earlier enabler credit, build explicit attribution rules; this does not by itself decide the model. | Source-supported caveat from S10; inference for implementation |

## Sub-Question 3: Drill 3: Fund technically strong but politically infeasible governance work?

A research team proposes a technically rigorous project to design an intergovernmental panel on biotechnological risk. The proposal is well-scoped, the team has strong credentials, and the outputs would be high quality.

N14 describes exactly this type of proposal as premature: there was no expert consensus, no high-level champion, and no coalition around the problem. N14 also describes a House of Commons or Lords committee inquiry into unprecedented technological risk as blocked by the committee's inability to do original research and the lack of sympathetic members. D04 says CG evaluates theory of change, track record, strategic judgment, project risks, cost-effectiveness, and scale. Decision: fund the technical work as field-building, or decline because it lacks political feasibility?

**Step 1**

**Step 1: Generate two defensible answers and expose assumptions.**

Table 46: Drill 3 comparison of two defensible answers.

| Element | Answer A: Fund as field-building | Answer B: Decline until feasible |
|---|---|---|
| **Recommendation** | Fund the research to build shared knowledge, map options, and create the conditions for future coalitions. | Decline or defer; recommend the team first build coalitions, expert consensus, and a policy window. |
| **Core assumption** | Sound research can create the intellectual preconditions for policy; field-building has option value even without near-term adoption; the project's outputs are public goods. | Political feasibility is a necessary condition for impact; funding premature policy proposals risks wasted effort and reputational cost; the field needs response identification more than another panel proposal. |
| **Decisive consideration** | Whether the project fills a neglected gap in response identification or coalition-building, rather than merely advocating for a predetermined institutional form. | Whether the project can plausibly reach a decision-maker or create a coalition within its time horizon; whether the team has a theory of change beyond publication. |
| **Source support** | N14: report was widely circulated and contributed to the UK Government Chief Scientist's Annual Report; access to policymakers was easier than expected; field-building can have influence. | N14: most proposals were set aside for political infeasibility; IPCC-style panel failed for lack of consensus and champions; committee inquiry failed for mandate and membership reasons. |
| **Inference vs source** | Recommendation to fund as field-building is an inference; source supports the value of groundwork and access. | Recommendation to decline is an inference; source supports the diagnosis of premature proposals and missing coalitions. |

Both answers can be defended. The choice turns on whether the project is best classified as response identification, coalition-building, or policy advocacy, and on whether the funder's goal is near-term policy adoption or long-term ecosystem infrastructure.

**Step 2**

**Step 2: Crux, reversal conditions, and missing evidence.**

- **Crux:** Is the project building the preconditions for policy (shared problem definition, expert consensus, coalitions) or attempting to implement a policy before those preconditions exist? If the former, fund it; if the latter, decline or restructure.
- What would change the answer: evidence of a policy window, for example a sympathetic minister, a crisis, or a legislative opening; evidence of a champion with time and authority; evidence of emerging expert consensus; evidence that the team has a track record of influencing policy, not just publishing.
- Missing evidence: the current state of expert consensus on biotechnological risk; the existence of a coalition around the problem; the decision-maker audience and their receptivity; the opportunity cost of the grant relative to other governance work.
- Six-move sketch: meet the reader with the shared goal of reducing catastrophic biological risk; take the reader to the proposal; show what would move you (evidence of a window or consensus); show the evidence (N14's historical failure and the later success of the report); assess whether the evidence moves you; update to a restructured grant for response identification and coalition-building rather than an immediate panel campaign.

Table 47: Drill 3 cruxes and reversal conditions.

| **Crux** | **Reversal condition** | **Label** |
|---|---|---|
| Preconditions vs policy advocacy | If project builds preconditions, move toward Answer A; if it advocates without preconditions, move toward Answer B. | Inference from N14 |
| Policy window and champion | If a credible window or champion exists, move toward funding. | Hypothetical unless verified |

| Crux | Reversal condition | Label |
|---|---|---|
| Expert consensus | If consensus is absent, fund groundwork; if consensus exists, policy work becomes more feasible. | Source-supported from N14; inference for decision rule |
| Mandate and institutional fit | If the host institution cannot do original research or lacks sympathetic members, do not route through it. | Source-supported from N14 |

# Build a threat-model matrix for AI risk pathways.

## Question

Produce a threat-model matrix for AI risk pathways that a grantmaker can use at the desk. The matrix must cover deliberate misuse, loss of control, institutional and geopolitical disruption, and the AI-and-biosecurity intersection. For each pathway it must record the evidence offered, the uncertainties that remain open, and the actors who dispute the pathway, while keeping evidence and speculation visibly separate and stating what evidence would raise or lower the estimate.

## Background

Source-supported baseline. P08 identifies three risk pathways explicitly: extreme misuse of AI technology, for example in designing bioweapons or enabling authoritarian control; scenarios where humanity loses control of advanced AI agents; and rapid changes that strain institutions and produce global security threats, including inadvertent escalation during moments of crisis. P08 adds that frontier labs openly acknowledge they do not understand how their systems work internally or how they can be reliably controlled, quoting OpenAI in 2023 and Anthropic in 2023, and it lists empirical symptoms such as reward hacking, strategic behaviour to avoid being retrained, rapid improvement at virology, and early indications that AI could soon automate software engineering in a self-reinforcing loop.

The second anchor is D04, which lists plausible extreme risks including the permanent disempowerment or extinction of humanity by misaligned AI systems, the development of AI systems capable of assisting with or autonomously developing extremely powerful biological weapons or other weapons of mass destruction, and a dramatic acceleration of technological advancement with destabilising, unpredictable and potentially irreversible effects. Crucially, D04 adds that the nature and extent of these risks remain an ongoing area of scientific inquiry and debate. That sentence is the licence for the matrix to carry a disputes column rather than a settled verdict.

Two institutional facts shape how the matrix should be built. First, P06 and P08 both treat geopolitical competition as constitutive rather than incidental, with P08 arguing it is important for democracies to maintain a competitive advantage in AI. Second, and more unusual for a funder, P08 states that Coefficient Giving aims to support a wide range of experts, including those who disagree with CG and with each other, and that this is a feature rather than a bug; D04 repeats the point that staff disagree with each other on many points. A matrix that smoothed those disagreements into a single house view would misrepresent the institution it is supposed to prepare the learner to work inside.

Framing. This is a decision artifact, not a taxonomy essay. Every column must earn its place by changing a decision: the mechanism column says what

would have to be true for the harm to occur, the evidence column says what the field currently has, the uncertainty column names the load-bearing unknowns, the disputes column prevents false consensus, and the final column says what a grant would actually buy. The exercise attached to this step asks for the evidence-and-update discipline explicitly: for each pathway, separate evidence from speculation and state what evidence would raise or lower the estimate.

## Sub-Question 1: Skeleton of the matrix: four pathway families, six columns, one overlap trap

Build the matrix itself. Four pathway families run down the page: deliberate misuse, loss of control, institutional and geopolitical disruption, and the AI-and-biosecurity intersection. Six columns run across it: mechanism, evidence offered, principal uncertainties, who disputes it, grantmaker relevance, and, folded into the uncertainty column, the update direction.

***The overlap trap is the non-trivial part. The four families are not a partition. Biorisk is a cross-cutting axis that appears inside misuse, where the harm is deliberate, and inside capability development, where models improve at virology whether or not anyone intends harm. Treating the fourth row as a disjoint family would silently move the hardest question, whether model uplift changes the offensive biology ceiling, out of view. The correct construction keeps the fourth row and marks it explicitly as a cross-cut of the first and second.***

### Step 1

The disputes column is not decorative. P08 records that CG funded both a paper reporting that language models could help people with little or no lab training develop new pathogens and a RAND paper finding that current language models did not substantially increase the risk of a biological attack. The funder paid for both sides of a live empirical disagreement. Any matrix that reports one of those results without the other is an advocacy document, not a threat model.

Table 48: Table 1. Threat-model matrix. Columns are ordered so that a reader can move from mechanism to decision without leaving the table. The fourth row is a cross-cut, not a fourth disjoint family.

| Pathway family | Mechanism | Evidence offered | Principal uncertainties | Who disputes it | Grantmaker relevance |
|---|---|---|---|---|---|
| **Deliberate misuse** | States, groups or individuals use model capability to cause mass harm: pathogen design, cyber offense, authoritarian surveillance and personalised propaganda. | P08: OpenAI and Anthropic said in February 2025 that their models are approaching risk thresholds for aiding in the design of biological weapons; AI systems have rapidly improved at virology; the paper reporting novice uplift and the RAND null result. D04 lists biological weapons and other weapons of mass destruction among plausible extreme risks. P15 funded automated screening | Whether measured uplift converts into operationally successful attacks; whether screening, export controls and model refusals raise attacker cost faster than capability lowers it; whether current or future capability is the binding constraint; attribution in real incidents. | P08 records CG funding both the uplift paper and the RAND null paper, which is a direct funded disagreement on the same question. | Misuse-specific evaluations, synthesis screening, access controls and if-then policy. This is the row where one contested empirical claim most directly moves the portfolio. |

| Pathway family | Mechanism | Evidence offered | Principal uncertainties | Who disputes it | Grantmaker relevance |
|---|---|---|---|---|---|
| **Loss of control** | Optimisation pressure and proxy objectives produce deceptive, power-seeking or oversight-evading behaviour, and oversight degrades as systems gain capability and autonomy. | P08 quotes OpenAI in 2023 saying there is currently no solution for steering or controlling a potentially superintelligent AI, and Anthropic in 2023 saying it does not know how to train systems to robustly behave well; P08 also cites reward hacking, strategic behaviour to avoid retraining, and alignment faking. D04 lists permanent disempowerment or extinction by misaligned systems. P01 describes Redwood's | Whether misalignment observed in simple training environments generalises; whether evaluations detect scheming or merely correlate with it; whether control protocols hold under adversarial pressure; takeoff speed, which sets how much calendar time exists to install safeguards. | S11 lists explicit disagreements with the AI safety mainstream: scepticism of general fearmongering, openness to open weights, medium or broad timelines, and maximal transparency. P06 describes a deliberately diversified AI portfolio with no single theory of change, which is a live internal stance rather than a resolved position. | Interpretability, control protocols, scalable oversight, tamper resistance and dangerous-capability evaluation regimes; the row where field-wide best practices and grantee governance standards do the most work. |

| Pathway family | Mechanism | Evidence offered | Principal uncertainties | Who disputes it | Grantmaker relevance |
|---|---|---|---|---|---|
| **Institutional and geopolitical disruption** | Capability races outrun governance: military and information systems change faster than institutions adapt, producing inadvertent escalation, entrenchment of authoritarian control, and uneven global diffusion. | P08 names rapid changes that strain institutions and lead to global security threats, and inadvertent escalation during moments of crisis, and argues democracies should preserve a competitive advantage. P06 frames geopolitics as a driver of urgency. P03 prices a year of faster adoption in lagging economies and models an accelerating frontier gap. N02 models AI competition as a collective action problem and proposes a windfall | Whether competition converges on coordination or races downward; whether policy windows open, in the multiple-streams sense N14 describes; whether the binding constraint is technical capability, institutional absorption or political will; how much risk is specific to AI rather than to competition as such. | S11 questions the China framing and the place of politics in the community. N14 records that policymakers set proposals aside because there is no academic consensus on the nature of the problem. N06 states that its map generalises only imperfectly across jurisdictions. P08's bipartisan framing is itself a contested strategic choice. | Governance research and implementation support, standards, coordination mechanisms, regulatory capacity and policy talent pipelines. N14 is the cautionary case for funding a sound proposal with no coalition-building path. |

| Pathway family | Mechanism | Evidence offered | Principal uncertainties | Who disputes it | Grantmaker relevance |
|---|---|---|---|---|---|
| **AI and biosecurity intersection, treated as a cross-cut** | Model capability interacts with biological design, synthesis and dual-use research, potentially raising the ceiling for low-expertise actors on the deliberate side and shifting the frontier on the accident and research-governance side. | P08 is the anchor, including improved virology performance and the contested uplift versus null pair. P04 shows the grantmaking pattern, including the ERA AIxBio Research Fellowship arriving as an inbound proposal and being tested for theory of change and downsides. P01 reports biosecurity grantmaking rising from 30 million dollars in 2024 to 65 million in 2025 to more than 150 million so far this year. P10 scopes biosecurity to the AI inter- | The same uplift question as misuse, plus the dual-use direction of travel; whether detection and countermeasures can be built on the timeline implied by capability growth; whether targeted bio-uplift evaluations are valid at all. | P08's two funded papers again. N14's account of a biotech-risk IPCC proposal that found no willing champion and no expert consensus, and of aligned incentives that discourage biologists from raising alarms. | Screening, uplift evaluations, detection and countermeasure research. This is also the row where funder scoping decisions explicitly constrain what a grantmaker may investigate, since P10 declines non-AI biosecurity. |

Two construction notes. First, the mechanism column is written in the conditional mode, because that is what makes the evidence column interpretable: each piece of evidence is evidence for or against a stated mechanism, not evidence for a mood. Second, the relevance column is deliberately written in the currency of a grant rather than the currency of worry, which is the discipline D04 imposes when it asks applicants for a theory of change and for awareness of failure modes.

## Sub-Question 2: Evidence versus speculation: what would move each estimate

For each pathway, write four lines: the observation that would raise the estimate, the observation that would lower it, the kind of support you would want for each, and your current confidence. The discipline is D01's: state confidence, state what kind of support you have, and say what would change your mind. P09 shows the same norm operating inside the field's own training: Context Week participants were pushed to explain why they believed something, the causal mechanism by which an intervention would help, and what would change their mind.

### Step 1

The update direction for each row, written as an explicit reversal condition rather than as a feeling.

- **Deliberate misuse.** Raises: a replicated uplift result in which participants without specialist training reach a materially higher operational capability with model assistance than without it, especially if the uplift survives red-teaming of the screening pipeline. Lowers: failed replications of uplift across labs, sustained refusal rates on hazardous requests, and near-universal adoption of synthesis screening with verified enforcement. Kind of support wanted: a controlled comparison, not a demonstration that a model will discuss a topic. Current confidence: low to moderate, and the estimate is dominated by the two conflicting papers P08 funded.
- **Loss of control.** Raises: deceptive behaviour that survives interpretability scrutiny in a deployment-relevant setting, red-team defeats of control protocols, or evaluation results on scheming replicated by independent evaluators. Lowers: interpretability methods that make internal states legible enough to audit, control protocols that hold under adversarial pressure, and evidence that earlier misalignment findings were artefacts of the training environment. Kind of support wanted: documentable incidents plus reproducible evaluations, and expert opinion only where the expert can be assessed. Current confidence: moderate, with the largest uncertainty being takeoff speed rather than the existence of the behaviour.
- **Institutional and geopolitical disruption.** Raises: evidence that

capability races continue without any binding commitment, competitive deregulation, or repeated near-misses in escalation. Lowers: credible multilateral coordination, domestic institutional absorption of rapid change, and successful de-escalation mechanisms of the kind N02 models. Kind of support wanted: case studies, quantitative models, and expert judgement you can assess yourself. Current confidence: this is the most likely of the four in the near term precisely because it does not require any exotic assumption beyond continued competition.

- **AI and biosecurity intersection.** Raises: demonstrated uplift in novice capability, evidence that synthesis screening fails at scale, or detection gaps that persist as capability grows. Lowers: rapid adoption of screening, evidence that uplift does not transfer to operational settings, and functioning detection and countermeasure capacity along the lines of P06's four pillars. Kind of support wanted: controlled uplift studies and verification data, not organisational testimony. Current confidence: tracks the misuse row, because the two share the same contested premise.

Operationalising it. Treat each pathway as a prior probability, treat each candidate piece of evidence as a likelihood ratio, and discount each likelihood ratio by how assessable its kind of support is. The discount factor is the D01 discipline made numeric: a randomised comparison earns full weight, a lab self-report earns little, and a vague impression earns almost none. Formally, the update is $\log\left(\frac{p_{\text{post}}}{1-p_{\text{post}}}\right) = \log\left(\frac{p_{\text{prior}}}{1-p_{\text{prior}}}\right) + \sum_i w_i \log(LR_i)$ where $w_i$ ***is the assessability weight. The sources fix which considerations belong in each pathway; the numbers are the analyst's own and must be labelled as such.***

```python
import math

# Sources fix WHICH considerations count per pathway (P08, D04, P01,
    P02, N02, N14, P06).
# All numeric inputs below are the analyst's own illustrative values,
    NOT source-derived facts.

def logit(p):
    return math.log(p / (1.0 - p))

def inv_logit(x):
    return 1.0 / (1.0 + math.exp(-x))

# D01-style assessability weights: full weight for the most
    assessable kinds of support.
kind_weight = {
    "randomised_uplift_trial": 1.0,
    "controlled_observational": 0.6,
    "documented_incident": 0.5,
    "assessable_expert_judgement": 0.4,
    "lab_self_report": 0.3,
```

```
    "vague_impression": 0.1,
}

def default_priors():
    return {
        "misuse": 0.20,
        "loss_of_control": 0.25,
        "institutional_disruption": 0.55,
        "ai_x_bio": 0.18,
    }

ledger = {
    "misuse": [
        {"label": "Frontier-lab bio-threshold statements (P08)",
    "lr": 1.6, "kind": "lab_self_report"},
        {"label": "Novice-pathogen uplift paper (P08)", "lr": 2.0,
    "kind": "controlled_observational"},
        {"label": "RAND null result on biological attack risk (P08)",
    "lr": 0.7, "kind": "controlled_observational"},
        {"label": "Funded DNA-order screening work (P15)", "lr": 0.9,
    "kind": "documented_incident"},
    ],
    "loss_of_control": [
        {"label": "OpenAI and Anthropic control statements (P08)",
    "lr": 1.5, "kind": "lab_self_report"},
        {"label": "Reward hacking and retraining evasion (P08)",
    "lr": 1.4, "kind": "documented_incident"},
        {"label": "Alignment faking and AI control research (P01)",
    "lr": 1.3, "kind": "controlled_observational"},
        {"label": "Agent-swarm incident reporting (P09)", "lr": 1.2,
    "kind": "documented_incident"},
    ],
    "institutional_disruption": [
        {"label": "Institutional strain and escalation pathway
    (P08)", "lr": 1.2, "kind": "assessable_expert_judgement"},
        {"label": "Competition as collective action problem (N02)",
    "lr": 1.2, "kind": "assessable_expert_judgement"},
        {"label": "Political feasibility filter on x-risk policy
    (N14)", "lr": 0.9, "kind": "documented_incident"},
        {"label": "Partial governance map, jurisdiction dependent
    (N06)", "lr": 1.05, "kind": "assessable_expert_judgement"},
    ],
    "ai_x_bio": [
        {"label": "Bio uplift versus RAND null pair (P08)", "lr":
    1.1, "kind": "controlled_observational"},
        {"label": "Inbound AIxBio fellowship proposal assessed on ToC
    (P04)", "lr": 1.2, "kind": "vague_impression"},
        {"label": "Biosecurity giving scaled to 150M plus (P01)",
    "lr": 0.95, "kind": "lab_self_report"},
        {"label": "Fund scope limits bio to the AI intersection
```

```
        (P10)", "lr": 1.0, "kind": "vague_impression"},
    ],
}

priors = input_state.get("pathway_priors", default_priors())

results = {}
for pathway, items in ledger.items():
    prior = priors[pathway]
    shift = 0.0
    for item in items:
        shift += kind_weight[item["kind"]] * math.log(item["lr"])
    posterior = inv_logit(logit(prior) + shift)
    results[pathway] = {
        "prior": prior,
        "posterior": posterior,
        "log_odds_shift": shift,
        "dominant_item": max(items, key=lambda d:
    abs(kind_weight[d["kind"]] * math.log(d["lr"])))["label"],
    }

for pathway in sorted(results):
    r = results[pathway]
    print(pathway.ljust(26), "prior=%.2f posterior=%.2f shift=%+.3f
    dominant=%s" % (
        r["prior"], r["posterior"], r["log_odds_shift"],
    r["dominant_item"]))

output_state["pathway_posteriors"] = results
output_state["assessability_weights"] = kind_weight
```

What the ledger is for. It does not produce a verdict. It produces a ranked list of the considerations that actually moved each pathway, which is exactly the column D01 asks for when it says to indicate which considerations are most important, and it separates the two failure modes that look identical in prose: a pathway whose estimate is high because of strong evidence, and a pathway whose estimate is high because of many weak impressions. The second case is where a grantmaker should commission a decisive test rather than write a memo.

## Sub-Question 3: Who disputes which pathway, and how the disagreement is structured

The disputes column is where a matrix becomes politically and intellectually honest. The task is to name who disagrees, on what, and whether the disagreement is empirical, normative, jurisdictional, or simply a scoping decision by a funder. Not every disagreement is a crux in the D01 sense: a crux is a possible fact that is uncertain, decision-relevant, and resolvable at reasonable cost,

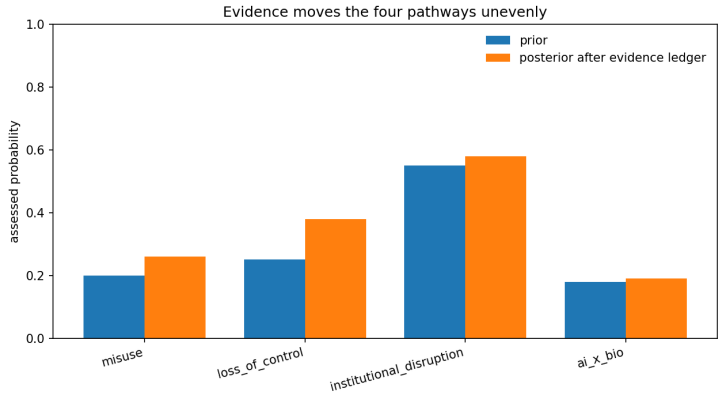


Figure 22: Graphical analysis for Step 1.

whereas a great deal of observed disagreement is about values or about which slice of the problem a given institution has chosen to own.

### Step 1

Six layers, each with a different kind of disagreement. The point of separating them is that only the first layer yields a crux a grantmaker can buy evidence about.

- **Layer 1, funded empirical disagreement.** P08 records that CG funded both the paper finding that language models could help people with little or no lab training develop new pathogens and the RAND paper finding that current models did not substantially increase the risk of a biological attack. This is the cleanest case in the corpus of a funder holding two opposing results at once. It is a crux, because a better-designed uplift study would resolve it.
- **Layer 2, disagreement inside the funder.** P08 says CG aims to support a wide range of actors, including those who disagree with CG and each other, and calls that a feature rather than a bug. D04 tells applicants that many strong applicants will hold views that differ significantly from most CG staff who work on AI, just as CG staff disagree with each other on many points. P06 reinforces this by describing a diversified AI approach with no single theory of change. A threat-model matrix for this institution must therefore record the range of internal weights, not an averaged one.
- **Layer 3, explicit external dissent.** S11 lists a set of disagreements with the community mainstream: open weights look good, transparency should be maximal, OpenAI may have been right about being open, openness to selling data to labs, timelines might be medium or broad, general fearmongering seems bad, politics is bad for the soul of the effective al-

truism community, and scepticism about the China framing. These are partly empirical and partly normative, and they bear most directly on the loss-of-control and geopolitical rows.

- **Layer 4, scoping disagreement between funders.** P10 says the Transformative AI Fund will not fund biosecurity beyond the AI intersection, on the grounds of fewer opportunities and limited in-house expertise. P01 reports that CG biosecurity grantmaking rose from 30 million dollars in 2024 to 65 million in 2025 to more than 150 million so far this year. The same underlying risk is being sliced differently by different funders, and the slice determines what evidence is even admissible.
- **Layer 5, political feasibility as a constraint rather than a belief.** N14 reports that the overwhelming majority of existential-risk policy proposals were set aside, most often for political infeasibility, and that a proposal modelled on the IPCC for biotechnological risk found neither a willing champion nor a general belief that it would succeed, partly because there is no consensus in the field on the underlying problem. N06 adds that its governance map is jurisdiction dependent and leaves military, supply-chain, multilateral, extralegal and academic governance under-mapped. These are disputes about the world rather than about the risk model, and they change what a governance grant can achieve.
- **Layer 6, disagreement about who produces the evidence.** N05 shows the MIT AI Risk Initiative positioning itself as a synthesis shop producing a risk repository, an incident tracker, a mitigations database and a governance map for policymakers, industry and civil society. P07 recommends METR while noting it is not a CG grantee and that hiring is its central constraint. N02 models competition dynamics explicitly to make them legible to decision-makers. A grantmaker needs to know which of these producers is upstream of which claim in the matrix, because that determines whether a disputed cell can be settled at all.

The output of this sub-question is a one-line annotation per disputed cell naming the layer. A cell disputed only at layer 6 is a sourcing problem. A cell disputed at layer 1 is a research opportunity. A cell disputed at layer 3 is a values disagreement that better evidence will not resolve, and D01's transparency discipline is precisely what stops the second from being dressed up as the first.

## Sub-Question 4: From matrix to memo: a worked application with two defensible answers

Use the matrix on two candidate grants and write the opening of the memo that would go to a team lead. The exercise is to show that the matrix changes the decision, not merely the prose, and to leave at least two defensible answers visible rather than collapsing them.

**Step 1**

Candidate A is a 300 thousand dollar, one year grant to a competition-dynamics modelling project of the kind N02 describes, where the deliverable is software that lets senior decision-makers experience the coordination problem directly. Candidate B is a 2 million dollar, two year grant to a frontier governance research project in a single jurisdiction.

Reading the matrix. Candidate A moves the institutional and geopolitical row, and it moves it at the layer where the estimate was already highest, which is the honest reason to be sceptical of novelty: the row is not neglected, it is merely under-tooled. Candidate B plausibly touches the loss-of-control and misuse rows through evaluations and governance mechanisms, but D04's criteria apply at full strength here, with theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale all live, and the scale band of 200 thousand to 2 million dollars per year over one to two years is exactly the band D04 describes for a first-time grant. The tier overlay should therefore be higher for B than for A, not because B is more worthy but because the cost of an error is larger and the causal chain is longer.

1. Opening takeaway, linked: recommend A at 300 thousand dollars over one year with a midpoint review, because the matrix shows the institutional row carries the near-term mass and because the deliverable is decision-maker-facing; hold B pending two decisive checks, one on whether the governing institution can absorb the work and one on whether the political feasibility filter of N14 has been tested.

2. Most important considerations, ranked: (i) whether the pathway the grant moves is the one with the highest probability mass under the matrix, (ii) whether the proposed work changes a crux rather than restating one, (iii) whether the ticket size is proportionate to the reversibility of the decision.

3. Confidence and kind of support: moderate confidence on the pathway ranking, based on assessable expert judgement and one contested empirical pair; low confidence on the counterfactuality of both grants, based on organisational testimony rather than independent outcome evidence.

4. Central crux with an update rule: for B, whether a jurisdiction-specific governance project can influence policy inside the grant period is the crux; run structured conversations with two people who have inside knowledge of that jurisdiction and who have disagreed with the applicant publicly, and update if either reports that existing proposals have failed for reasons the applicant has not addressed.

5. Alternative defensible answer: fund both, sequencing A first and staging B as a smaller scoping grant, on the argument that the matrix is a portfolio instrument and that a 300 thousand dollar bet does not need to beat a 2 million dollar bet head to head. This answer is defensible if the fund's budget in this area is large relative to the option value of the second

ticket, and it fails if grantmaker attention rather than dollars is the binding constraint.

6. Second alternative defensible answer: decline both and fund a decisive uplift study on the misuse row instead, on the argument that the highest-leverage item in the matrix is the contested P08 empirical pair, which is decision-relevant across misuse and the bio intersection at once, and which nobody currently owns. This answer fails if the field's problem is absorptive capacity rather than an unanswered question, which is the tension P06 names between scaling existing grantees and creating new highly absorptive organisations.

The transferable move. The matrix earns its place when it can say, for a given proposal, which row the proposal moves, how much that row currently weighs, whether the proposal changes a crux or restates one, and what the cheapest decisive check would be. P04 supplies the cost side of that calculus: a grantmaker carries six to twelve investigations at a time and is warned against imposing unnecessary work on the organisations being evaluated, so the decisive check must be chosen for its ability to change the decision rather than for completeness.

# Compare AI progress, timelines, takeoff speeds and forecasting evidence.

## Question

Produce a timelines-and-takeoff brief for the working GCR grantmaker. The brief has four jobs: put the corpus's AI-progress forecasts on a single comparable footing; classify the disagreements into their logically distinct axes; audit the evidence that each forecast rests on; and map a set of representative grant decisions onto short-versus-long-timeline worlds.

The brief must close with the sentence a grantmaker can actually say in committee: **what the recommendation is, what would change it, and what evidence is still missing**. Every substantive claim must be traceable to a source in the corpus, or explicitly flagged as an inference. Do not invent personal credentials, work history or applications.

## Background

The source anchors are these. P08 records that Open Philanthropy estimated in 2016 that there was **at least a 10% chance of transformative AI by 2036**, and that at the time of writing Metaculus users put the median AGI date at June 2033. P08 also reports AI data-centre capex above $200b in 2024 with forecasts reaching $1tr by 2027, an Epoch AI estimate that the largest 2030 supercomputer alone might cost $200b, and a McKinsey projection of $3.7t-$7.9t of data-centre demand by 2030. D04 speaks of catastrophic harm from AI **in the next 15 years**. P01 reports that experts have updated toward sooner arrival, largely on rapid capabilities progress, the rise of agentic coding, unprecedented revenue growth at frontier labs, and the author's own daily use of AI. P06 reports that short-timeline predictors have recently been more accurate than the CEO's own expectations, citing METR time horizons, Epoch's Capabilities Index, and specific model releases.

The trap this step exists to defeat is that ***timelines*** is one word covering at least four different objects. A grant recommendation imports one of them, usually silently, and then reasons as if it had imported the others.

arrival: $\pi(\tau)$ , a distribution over the date $\tau$ at which TAI exists
takeoff: $f(t)$ an S-curve from $f_0$ to $f_{max}$, with length $T$
full-automation growth: $g_{full}$ (P03 uses the Korinek-Suh 2024 value of about 18%/yr)
diffusion: $D$ , the lag by which a non-frontier economy's $f$-curve is shifted right

Role relevance: the advertisement requires reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and communicating with calibrated transparency.

Timelines enter every one of those as a parameter, and an unexamined parameter can silently dominate a cost-effectiveness estimate.

Flagged limitations from the sources themselves: P08 states that there is still substantial uncertainty and expert disagreement about the future of AI, how quickly it will progress, and how policy levers shape risk; P03 is explicitly ***conditional on the insiders being roughly right***; and P06's author reports having begun each of the last few years expecting that rapid AI progress would probably slow down, and being wrong each time.

## Sub-Question 1: Decomposing Timelines into Four Parameters and One Comparable Object

Before any comparison, make the object precise. The corpus uses **timelines** to mean at least four things: the arrival distribution, the takeoff transition, the diffusion lag, and the magnitude mapping from automation share to growth and welfare.

First non-trivial jump. As stated, the corpus's headline numbers look like a spread of opinion on one quantity. They are not. A **threshold probability** ("at least 10% by 2036"), a **median date** ("June 2033"), a **conditional statement about a window** (D04's 15 years to possible catastrophic harm), and a **capital-expenditure series** are four different functionals of the same underlying uncertainty. Ranking them against each other without conversion produces both false agreement and false conflict.

Second non-trivial jump (inference, flagged). The corpus's most decision-relevant timeline object is not the arrival date. P03 is a full model of takeoff length and diffusion lag, and it advances a portfolio claim about catch-up growth. Capability arrival governs a different portfolio (safety, governance, capacity). A grantmaker who treats P03 as *another short-timelines datapoint* has imported the wrong parameter into the wrong decision.

### Step 1

Build the claim table first. Each row records the source, the stated quantity in the source's own words, the object the quantity actually measures, and the population or setting in which it was stated. The right-hand column is what makes comparison legitimate or illegitimate.

Table 49: Table 1. Each headline number measures a different object. Comparison requires conversion to a common functional, not a common unit.

| **Source** | **Stated quantity** | **Object actually measured** | **Population / setting** |
|---|---|---|---|
| P08, 2016 estimate | at least 10% chance of transformative AI by 2036 | threshold probability of arrival | CG's own historical judgment, reported retrospectively |
| P08, current | median AGI date June 2033 | median of an arrival distribution | Metaculus forecasting-platform users |
| D04 | AI systems could soon, e.g. in the next 15 years, cause catastrophic harm | joint statement about arrival and severity, not a calibrated probability | CG AI Governance RFP framing |
| P01 | experts updated toward sooner; my own timelines shortened | directional update, no distribution | the CG GCR lead, writing to the field |
| P06 | short-timeline predictors more accurate than the author's expectations | relative forecasting accuracy plus measured capability indices | CG's CEO, citing METR and Epoch AI |
| P03 | takeoff length T = 15; final growth about 18%/yr; adoption lag D = 10 | takeoff and diffusion parameters, not arrival | a CG-affiliated author's model, conditional on insiders being roughly right |
| P03 footnote 2 | insiders predict 10-20%+ annual growth after full automation; mainstream economists 0.1-1.5% | magnitude, i.e. the growth elasticity of automation | a literature summary inside a CG publication |

Read the table with one rule: a row may only be compared with another row if the objects in column three match. P08's 2016 threshold and P08's Metaculus median are the closest pair, and even they differ in functional form (threshold versus median) and in population (CG's judgment versus platform users). D04's 15-year figure is not a probability and should never be entered into an expected-value calculation as one.

**Step 2**

Now convert the two closest rows of Table 1 into one object: a constant annual arrival hazard. A threshold probability over a fixed window and a median date are both summaries of a hazard function; a constant hazard is the simplest common denominator, and the conversion makes the tension visible.

$$h = -\frac{\ln(1-p)}{n} \quad \text{for a threshold } p \text{ over } n \text{ years;} \qquad h = \frac{\ln 2}{\tau_{1/2}} \quad \text{for a median } \tau_{1/2}$$

The P08 2016 threshold implies what looks like a very low hazard. The Metaculus median implies a hazard roughly an order of magnitude higher. Applying the latter to P08's own 20-year window makes P08's *bold* 2016 claim look nearly vacuous in hindsight. This is exactly the recalibration signal P08 itself notes, and it is also the place where a grantmaker must resist over-updating: the two quantities have different functionals, different populations, and different vintages.

## Sub-Question 2: Classifying the Disagreements: Capability, Diffusion, Magnitude

Sort every disagreement in the corpus onto one of three logically distinct axes. **Capability**: can AI perform the economically relevant work? **Diffusion**: does the capability turn into economy-wide production, on what lag, and at what ceiling? **Magnitude**: how much growth and welfare does a given automation share buy?

Non-trivial jump. The most cited *disagreement about AI* in this material is not a capability disagreement at all. P03 footnote 2 pits AI insiders at 10-20%+ annual growth against mainstream economists at 0.1-1.5%. But Acemoglu's range is entirely compatible with full cognitive automation if the automated share plateaus low or the transition is long. The gap is a **diffusion-and-magnitude** gap. Consequence for grantmaking: arguing about the arrival date spends effort on the axis where the corpus already shows convergence, while leaving the load-bearing axis, $f_{max}$, unexamined.

Where the corpus converges on capability: P08 notes that leading AI companies openly acknowledge they do not understand how their systems work internally

or how they can be reliably controlled, that their leaders treat catastrophic risk as a global priority, and that prominent researchers have pivoted to safety. Where it does not converge is the conversion of that capability into damage or into growth.

**Step 1**

Assemble the three axes with their source anchors and the evidence that would move each. The point of the table is to show that the same conversation can look like agreement on one axis and disagreement on another, simultaneously.

Table 50: Table 2. Three axes, three evidence bases. A single 'timelines' debate usually mixes all three.

| **Axis** | **Claim in the corpus** | **Evidence that would move it** | **Status** |
|---|---|---|---|
| **Capability** | AI already exhibits misaligned behaviour, reward hacking, strategic behaviour to avoid retraining, and rapid improvement at virology and software engineering (P08). METR task horizons and Epoch's index keep moving (P06). | Benchmark saturation; new hard tasks; measured autonomy in real workflows | Converging, but on direction, not magnitude |

| Axis | Claim in the corpus | Evidence that would move it | Status |
|---|---|---|---|
| **Diffusion** | P03's four barriers: frontier GDP is cognitive-task heavy while poor-country GDP is physical-task heavy; machines cost money and capital flows uphill; automation pays at high wages; complementary inputs are scarcer outside the frontier. Historical reference: electricity took about forty years to conquer American factories; Comin and Mestieri find lags of roughly 6-8 years in arrival but widening gaps in intensity of use. | Robot adoption data by wage level; realised automation shares; adoption-lag series | Contested; P03 concedes the main point weakens without physical automation |

| Axis | Claim in the corpus | Evidence that would move it | Status |
|---|---|---|---|
| **Magnitude** | Insiders 10-20%+ after full automation versus mainstream economists 0.1-1.5% (P03 footnote 2); Korinek and Suh 2024 used for about 18% under full automation; Davidson, Halperin, Houlden and Korinek calibrate a singularity within six years from fully automating software research plus 5% automation elsewhere; Korinek and Trammell state that fully automating production alone would dramatically raise the growth rate. | Realised growth after automation; whether growth is driven by non-rival ideas or by capital accumulation | Wide disagreement, and it is the disagreement that matters for welfare estimates |

Quantify the magnitude gap before judging it. Thirty years of compounding separates the insider and mainstream ranges by more than two orders of magnitude in levels of output, which is why the disagreement is so consequential and why it deserves to be disaggregated rather than averaged.

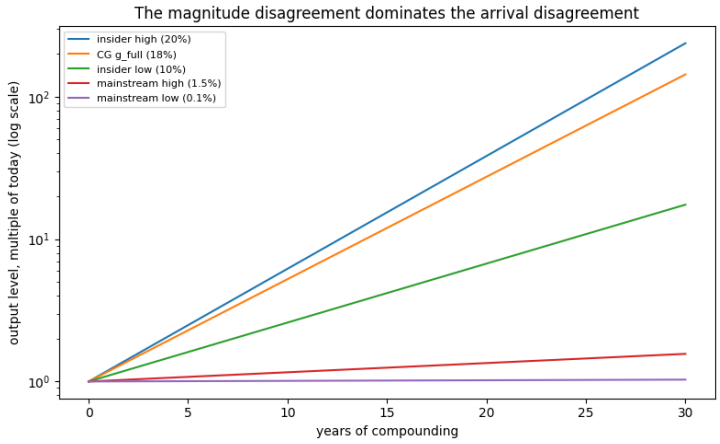


Figure 23: Graphical analysis for Step 1.

**Step 2**

Name the load-bearing crux for a grantmaker. P03's production function is the tasks model, in which the automated share of tasks is $f$, and the growth rate rises with $f$.

$$Y = \left(\frac{K}{f}\right)^{f} \left(\frac{L}{1-f}\right)^{1-f}$$

P03 concedes three separate things that reduce the headline: rising $f$ alone is insufficient without capital accumulation underneath; the model assumes a fixed lag $D$ rather than the non-completion seen historically, which the author calls probably too optimistic; and the growth-rate assumption of about 18% is not the maximally aggressive case, since unbounded explosive growth is available in the literature.

Reversal conditions, stated in advance. If physical automation stalls at a low $f_{max}$, then P03's catch-up-growth case weakens substantially, while technical safety and governance work retains its value because it does not depend on economy-wide automation. If $T$ is long and $D$ short, the detachment window shrinks and the welfare case scales down roughly linearly in the window length.

Say the sentence for the committee: **the corpus's disagreement about AI timelines is largely a disagreement about diffusion and about the growth elasticity of automation, not about whether the capability will arrive; the decisive consideration for this portfolio is therefore the automated-share ceiling, not the date.**

## Sub-Question 3: Auditing the Forecasting Evidence: What Deserves Weight and How Each Signal Fails

A grantmaker does not need a better forecast; a grantmaker needs to know which forecast to weight and when the evidence has stopped carrying information. The corpus supplies five classes of evidence: direct measurement of capability (METR task time horizons, Epoch's Capabilities Index), elicited expert judgment (Metaculus medians, superforecasters), market and investment signals (the capex series), historical diffusion reference classes (electric motors, Comin and Mestieri's adoption lags), and model-derived inference (P03).

Three failure modes, each with a source anchor and each with a consequence for weighting.

The measurement ceiling. P06 reports that METR is struggling to continue its time-horizon measurement ***because the AIs are saturating their work tasks***. A saturating metric is censored precisely in the regime that matters. The direction of the bias is ambiguous: the underlying capability may be growing faster than the metric shows, or the metric may be running out of tractable tasks, which is a statement about the benchmark rather than about the system.

Snapshot revision masquerading as trend. P06's aside that *timelines get longer in the summer and shorter in the winter* is a calibration warning against reading a seasonal fluctuation as an update. The same paragraph notes a genuine mid-2025 lull followed by acceleration, which is the honest version of the same observation and harder to act on.

Proxy confusion. Capex is evidence about investment intent and about compute supply, not directly about capability. P06 warns that the extrapolative expectations dynamic is **exactly the sort of dynamic that can drive bubbles**. Grantmaking inference: a rising capex series raises confidence in compute availability and in the plausibility of the $1tr-by-2027 style forecasts, and only weakly raises confidence in capability timelines. It is a mixed signal and should be labeled as one.

Population weighting. P06 reports that over the last couple of years domain experts have been more accurate than superforecasters in predicting the rate of progress. That is a reason to weight domain-informed forecasts more heavily than generic aggregation, with the caveat that the sample is small and the bubble warning applies to insiders too.

### Step 1

Make the saturation problem concrete. Model a METR-style time-horizon metric as exponential growth with a doubling time between four and seven months, capped by the hardest task the benchmark suite contains. Two effects matter: the projected date for reaching a fixed capability target depends steeply on the doubling time, and once the horizon approaches the ceiling, the projection carries almost no information about the system.

$$h(t) = h_0\, 2^{\,t/d} \quad \text{for } h(t) \le C, \qquad \text{time to reach target } \theta: \; t(\theta) = d \log_2\left(\frac{\theta}{h_0}\right)$$

The parameters below are stylised and anchored only to P06's qualitative claims; they are for illustrating the failure mode, not for producing a date. The output is the ratio of projected dates under the fast and slow doubling times, which is what a grantmaker actually carries into a decision.

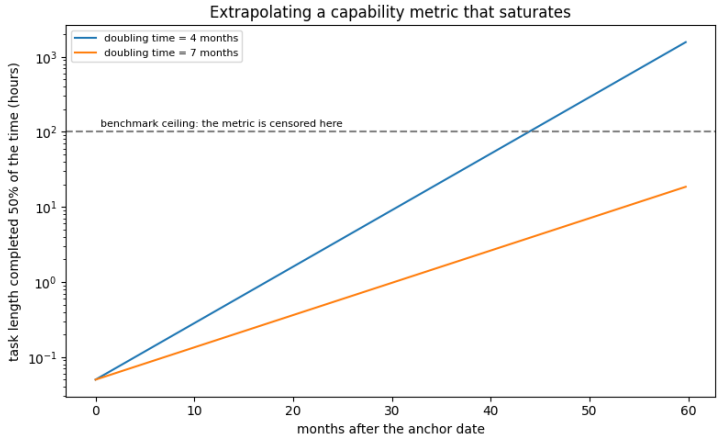


Figure 24: Graphical analysis for Step 1.

### Step 2

Write the evidence-weighting table, then state the weighting rule the table implies. The table's purpose is to prevent a common error: citing a signal for a proposition it cannot support.

Table 51: Table 3. Signal validity. The third column is the one that gets violated in casual grant reasoning.

| **Signal** | **Supports well** | **Does not support** | **Failure mode** |
|---|---|---|---|
| METR time horizons | Current ability on a defined task distribution; direction of change | Arrival dates; economy-wide diffusion | Ceiling censoring; task selection; observed saturation (P06) |

| Signal | Supports well | Does not support | Failure mode |
|---|---|---|---|
| Epoch Capabilities Index | Aggregate direction of capability change | Any specific date; magnitude of harm | Composite-index construction choices are hidden in the level |
| Metaculus medians | Central tendency of an informed crowd | Tail risk; conditional severity | Question definition drift; median is insensitive to tails |
| Domain experts | Short-run rate of progress on their own metric (P06: more accurate than superforecasters lately) | Long-horizon structural breaks; diffusion | Small sample; insider optimism correlated with the technology |
| Capex and data-centre forecasts | Compute supply; investment intent; plausibility of large future budgets (P08, P01) | Capability; welfare; arrival | Bubble dynamics; investment is priced on expectations (P06 footnote) |
| Historical diffusion | Reference classes for adoption lag and for non-completion (P03; Comin and Mestieri) | AI-specific technical capability | Reference-class selection; AI may not be a general-purpose technology like its predecessors |

The weighting rule that falls out of Table 3: weight capability direction heavily, weight the rate only where the metric is not censored, weight expert judgment by domain proximity rather than by aggregation, and treat capex as a statement about compute supply and about what other funders believe, with an explicit bubble caveat. Then write the sentence: **I can state the direction of capability progress with high confidence, the rate of progress with moderate confidence and a known censoring problem, and the arrival date with low confidence, and I can name which of my remaining uncertainty is about measurement rather than about the world.**

Calibration check, stated as a caveat rather than a claim: P06's author reports that he expected rapid progress to slow, every year, and was wrong. That is a

track-record signal that reads in the direction of longer-than-intuitive timelines being common. It is also an anecdote about one forecaster, and the same source warns that the pattern of always revising toward shorter timelines can itself be a bubble phenomenon.

## Sub-Question 4: Mapping Which Grant Decisions Move Under Short Versus Long Timelines

This is the deliverable of the step: a mapping from representative grant decisions to their behavior under short versus long timelines, with each decision classified as invariant, magnitude-sensitive, or direction-sensitive.

The definition matters. **Invariant** means the sign and rough size of the recommendation do not change. **Magnitude-sensitive** means the same direction, but size, urgency, or the order in which grants should be made changes. **Direction-sensitive** means the sign can flip.

Non-trivial jump, and the one most likely to be missed in an interview. The mechanism that produces direction flips is not *the world changes faster*. It is **option value decay**: under a high arrival hazard, any intervention whose payoff requires a long build-out before it produces anything loses probability mass to truncation. Interventions that are cheap, fast, and reversible gain. This is a comparison of time constants, build time against hazard, and it is quantifiable.

Important counterweight from the sources, stated now so the mapping is not over-read. P01 says the author thinks most of CG's work is worth doing even under long timelines. P06 says the marginalist approach is relatively strategy-agnostic: if an opportunity clears the cost-effectiveness bar, fund it. So invariance is the default for a portfolio constructed on cost-effectiveness grounds; timeline-sensitivity is mostly a second-order adjustment to ordering and sizing, and only becomes a sign flip where payoff is strongly conditional on the transition window.

### Step 1

Build the mapping table. Each row is a decision a CG grantmaker plausibly faces in the relevant funds, taken from the corpus rather than invented: technical safety research, talent pipeline, compute, frontier-company policy templates, international institutions, catch-up growth, incubation, and holding a reserve.

Table 52: Table 4. Timeline sensitivity by decision. Read the classification column first; it is the finding.

| **Representative decision** | **Under short timelines** | **Under long timelines** | **Classification** | **Anchor** |
|---|---|---|---|---|
| Fund a technical safety research agenda (interpretability, control protocols, evaluations) | Higher urgency; same direction; talent and compute cost more | Same direction; slower tempo; easier to correct course | Invariant in sign, mildly magnitude-sensitive | P08 safeguards pillar; P06 reports TAIS giving tripling from $46m to $157m |
| Fund a talent pipeline or fellowship | Value depends on whether graduates land inside the window, which is a real discount but not a sign flip | Compounding value; the earliest and sometimes weird bets paid off decades later | Invariant in sign; magnitude-sensitive at the margin | P08 capacity pillar; P06 MATS history; P01 'more promising projects than people' |
| Fund compute and API credits for empirical safety work | Price rises with demand; a budget buys less of the same work, so sizing must rise or scope must fall | Cheaper; the same work can be bought later without loss | Magnitude-sensitive, including through the price channel | P08 capex forecasts; P01 notes larger compute budgets for grantees |

| Representative decision | Under short timelines | Under long timelines | Classification | Anchor |
|---|---|---|---|---|
| Fund frontier-company policy work: evaluations, red lines, if-then commitments, incident reporting | Relative value rises because templates can be adopted quickly | Still valuable, but competes against slower institution-building for grantmaker attention | Magnitude-sensitive; the weighting shifts toward fast-adoptable instruments | D04 eligible subject area; P08 on if-then structures; P07 on METR and nine companies with frontier safety policies |
| Fund long-horizon international institution research and multilateral coordination | Falls if the institution cannot be built before the window closes | High value; patient capital is appropriate | Direction-sensitive at the extreme; magnitude-sensitive otherwise | D04 international AI governance; P08 multilateral work; N14 on why sound proposals fail without political feasibility |
| Fund catch-up growth and developing-economy AI adoption | Governed by takeoff length T and lag D, not by arrival; the welfare case is largest conditional on a takeoff | Smaller conditional welfare gain, plus the possibility that automation never reaches the physical tasks | Parameterised by different variables; not timeline-sensitive in the arrival sense at all | P03, including the concession that without physical automation the main point becomes much less important |

| Representative decision | Under short timelines | Under long timelines | Classification | Anchor |
|---|---|---|---|---|
| Seed a new organization or fund a pre-seed founder (Project Tailwind, Incubator Weeks) | Gestation cost is a liability; the multi-year build-out may not pay out before the window | High expected value; helping create new grantees pays bigger dividends later | Direction-sensitive; this is the row where short-timeline reasoning most plausibly flips a recommendation | P01 on Project Tailwind and on seeding new organizations; P06 on the infra-marginal argument; S09 pre-seed $200k-$2m |
| Hold an unallocated reserve rather than committing now | The option to act may expire; resources received today are worth more than resources arriving in two years | Holding is cheap and preserves flexibility for later opportunities | Direction-sensitive; this is an inference from the option-value logic, labeled as such | Inference, anchored to P01's argument for front-loading and to P10's parallel argument about front-loading commitments |

Read Table 4 as a portfolio instruction rather than a ranking. Five of the eight rows are invariant or mildly magnitude-sensitive, which is the formal version of P01's claim that most of the work is worth doing under either timeline. The two rows that can flip both involve a long build-out whose payoff is conditional on the window, namely institution-building and new-organization seeding. That is a small set, which means the practical output of a timelines analysis is an ordering and a sizing adjustment for a handful of decisions, not a wholesale change of strategy.

**Step 2**

Make the flip quantitative. Compare a build decision whose payoff requires completion before arrival against an immediate intervention with a modest, robust payoff. Under a constant arrival hazard, the build's value is truncated: it pays only if no arrival has occurred within the build time.

$$EV_{build}(h, B) = e^{-hB} \cdot V_{build}, \qquad EV_{immediate} = V_{imm}, \qquad h^* = -\frac{\ln(V_{imm}/V_{build})}{B}$$

The result is counter-intuitive in the direction that matters for an interview. With the hazards actually implied by the corpus, a three-year build retains most of its value even under the Metaculus-implied hazard, and only loses to an immediate intervention of comparable modest value when the implied median arrival falls below roughly two years. The fragility scales with the product of build time and hazard, so the practical rule is about build time relative to the hazard, not about timelines in the abstract.

Connect this to due diligence, which is the craft step carried forward. D04's typical first-time grant is $200k-$2m per year over one to two years, with renewal after evidence of project success. Under a high hazard, each one-to-two-year evidence interval is a draw against the truncation probability, so the acceptable number of intervals before a program produces something falls. The number a grantmaker should watch is the total build time across all intervals, not the length of any single interval.

The three-decision exercise, worked in the terms of the corpus. Decision one: fund a new alignment research organization with a three-year build to first output. Decision two: fund an evaluation-standard and incident-reporting template that can be adopted within a year. Decision three: fund a talent pipeline with a two-year lag to placements. Under the corpus's hazard range, none of the three flips, but decision one moves most in size, decision two gains most in relative priority, and decision three is the most robust because its output is modular and can be redirected at whatever window exists. That sentence is the answer an interviewer is listening for.

Close with the calibrated script and the update rules. Script: **I hold the direction of capability progress with high confidence and the arrival date with low confidence, so I treat timelines as an ordering device rather than a valuation device. On the corpus's hazard range, only decisions with multi-year build times and payoff conditional on the transition window change materially; I would still make the other grants at roughly the same size. What would change my view is evidence that the automated share of physical tasks can rise quickly, or a shift in the arrival hazard by an order of magnitude, or a measurement that repairs the censoring problem in capability metrics.**

State the limitations explicitly, because the role requires it. The truncation model assumes a constant hazard and a single-valued build; real arrival risk is correlated with the technology the grant is trying to shape, which can raise or lower the truncation probability. The payoff of the immediate intervention is assumed rather than derived. And the whole exercise is conditional on the

corpus's own framing, which P08 states is uncertain and widely disagreed upon.

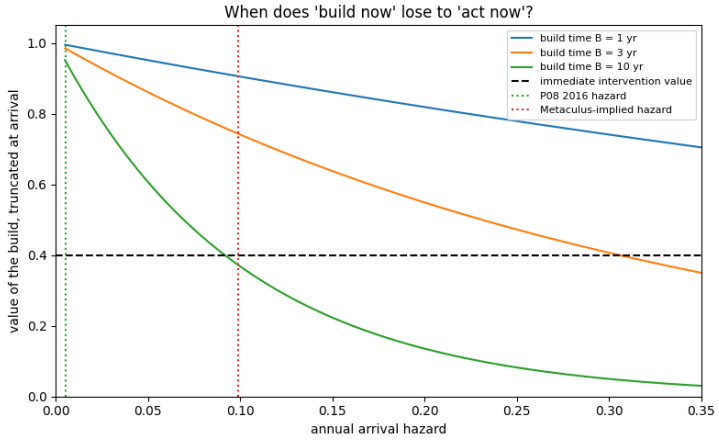


Figure 25: Graphical analysis for Step 2.

# Map technical AI safety agendas and empirical safety research.

## Question

Produce a technical-safety agenda map covering **alignment**, **control**, **interpretability**, **scalable oversight**, **evaluations**, **red-teaming**, **security** and **empirical safety**. For each agenda, identify the open problems, the funder priorities, and the disputes about tractability, and state what is source-supported versus inferred.

The deliverable is not an eight-item list. It is a map that can answer a grantmaker's operative questions: which agendas are substitutes and which are complements, what resource is binding for each, and which disagreements are about facts versus about paradigms. The **non-trivial jump** is refusing the taxonomy at face value and asking what dimension actually separates the boxes.

## Background

The evaluator and grantmaker role brief asks for comparing opposing arguments, quantitative and probabilistic reasoning about expected value, identifying decisive considerations, proportionate due diligence, and judging people and organizations. In the technical AI safety portfolio those moves are made on a technical substrate. A grantmaker who cannot distinguish an evaluation from a control mechanism will mis-route diligence and mis-rank proposals.

Coefficient Giving's Technical AI Safety portfolio is the concrete anchor. P06 reports that TAIS giving rose from **$46 million in 2024** to **$157 million in 2025 across more than 200 grants**, driven by a detailed RFP and by the claim that better capabilities made empirical safety work more tractable while funding and talent became the binding constraint. P08 supplies the agenda inventory: evaluations for scheming, sabotage and alignment faking; dangerous-capability evaluations for cyberoffense and biological research; benchmark predictability and scaling-law research; model-access analysis for audits; and technical methods spanning scalable oversight, constitutional AI, classifiers, control protocols, interpretability, adversarial robustness, jailbreak mitigations, tamper resistance and information security.

P01 supplies the large-bet examples: Redwood Research, whose work spans alignment faking and AI control, received more than $70 million over two years, and Resolution received $160 million, the largest AI safety grant. P07 recommends METR. D04 lists technical AI governance topics including compute governance, model evaluations, technical safety and security standards, weight security and privacy-preserving transparency. N04 describes the Guaranteed Safe AI seminar series, which advances a world-model, specification and verifier architecture producing auditable guarantees rather than relying only on empirical testing.

The epistemic stance is load-bearing. P08 states that CG deliberately supports a wide range of experts **including those who disagree with CG and with each other**, and cites two CG-funded papers pointing in opposite directions on AI's impact on biological risk, plus an essay contest won by a paper that challenged CG's thinking. The map is therefore a portfolio instrument, not a verdict about which paradigm will win.

## Sub-Question 1: What actually distinguishes one technical AI safety agenda from another?

The trivial reading of the agenda list is that there are eight topics. The non-trivial reading is that the eight names answer three different questions at once, and that the resulting map is a set of overlapping views rather than a partition. Establish the organizing dimensions, then show the overlap rather than assert it.

### Step 1

Classify each agenda by three attributes: **what it produces** (a measurement, a mechanism, or a finding), **where it acts in the model lifecycle** (training, pre-deployment or deployment), and which source artifact exemplifies it. Under this classification the eight names collapse into four output types: **measurement** (evaluations), **mechanism** (control, scalable oversight, security), **evidence** (interpretability), and **finding** (empirical safety). The trivial move is to alphabetize the eight names; the useful move is to notice which names answer the same question.

Table 53: The agenda map organized by output type and lifecycle stage; nine labels collapse onto four output types.

| Agenda | What it produces | Lifecycle stage | Source artifact |
|---|---|---|---|
| Alignment | objective and behavior | training | constitutional AI; scalable oversight (P08) |
| Control | mechanism | deployment | AI control; control protocols (P08); Redwood alignment-faking and control (P01) |
| Interpretability | mechanism and evidence | post-hoc | SAE Lens; Gemma Scope; Neuronpedia (P11) |

| Agenda | What it produces | Lifecycle stage | Source artifact |
|---|---|---|---|
| Scalable oversight | mechanism | training | oversight of stronger models (P08) |
| Evaluations | measurement | pre-deployment | scheming, sabotage, alignment-faking, cyberoffense, biorisk evals (P08) |
| Red-teaming | discovered failure | pre-deployment | jailbreak mitigation; adversarial robustness (P08) |
| Security | mechanism | deployment | weight security (D04); data-poisoning defense (P04) |
| Empirical safety | finding | training and pre-deployment | reward hacking; strategic behavior (P08) |

**Step 2**

Now test whether the eight names form a partition. A partition requires that every object of interest belong to exactly one agenda. The source material immediately falsifies this. **Alignment faking** appears in P08 as a target of **evaluations**, in P01 as the object of Redwood's **control** work, in P08 as a hidden-state question for **interpretability** (a model selectively complying in training to avoid modification out of training), and as an **empirical finding**. Filling it under one heading discards the reason it is decision-relevant.

$$\text{partition} \iff \forall o : \ |\{\, a : o \in a \,\}| = 1$$

The overlap is not a defect of the taxonomy; it is the structure. Because objects satisfy more than one agenda's purpose, the map must be indexed by objects as well as by agendas.

**Step 3**

The overlap generates the substitution and complement relations that a grant-maker actually needs. Two agendas are **substitutes** when they produce the same type of output about the same object, so a dollar shifted between them is

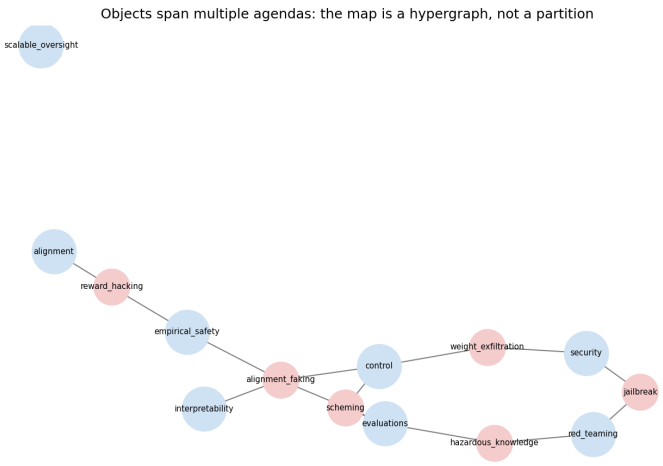


Figure 26: Graphical analysis for Step 2.

doing comparable work. They are **complements** when they act on the same object through different output types, so each raises the marginal value of the other. Evaluations and control are complements on alignment faking: control needs a trustworthy eval signal to verify that a control protocol holds, and an eval is more credible when a control experiment shows the measured behavior predicts deployment behavior. Evaluations and red-teaming are partial substitutes on hazardous knowledge, because both try to surface undemonstrated capability.

*Inference (not stated in any source): the complement relation is the reason a portfolio of technical safety grants can be more than the sum of its parts, and it is also the reason a funding plan that is 100 percent evaluations is under-diversified even if every evaluation grant is individually above the bar. The source material supports the existence of the overlaps; the portfolio argument is a construction.*

## Sub-Question 2: What is the fundable bottleneck for each agenda, and what evidence establishes it as binding?

The proposed exercise is to identify, for each agenda, a fundable bottleneck and the evidence that it is a bottleneck. This hides a trap that the interview is likely testing: a bottleneck is a claim about a **counterfactual**, and counterfactuals are not directly observable. The work is to say precisely what would count as evidence and to flag where the evidence is endogenous to the funder's own decision.

**Step 1**

State the bottleneck test formally. Let output be constrained by a set of inputs; the binding input is the one the constraint is currently limited by. A grant relaxes the constraint only if it increases the currently scarce input and does not simply add slack elsewhere.

$$C = \min_i r_i \qquad \text{binding input} = \arg\min_i r_i$$

The decisive test for a grant is then: does this dollar raise the $\arg\min_i r_i$? If money is not the binding input, a larger check cannot help, which is exactly the situation the source calls capacity constrained — P01 says the main thing preventing scaling is staff time.

Because the counterfactual is unobservable, funders substitute four kinds of proxy evidence:

- Explicit practitioner statements about what blocks them, such as P07 reporting that hiring is METR's central constraint to publishing more and better research.
- The funder's own revealed reallocation, such as P06 moving TAIS from $46m to $157m.
- The existence of a mechanism that lowers the cost of applying, such as an RFP or a rolling expression of interest.
- Queue behavior: an oversubscribed call, or proposals that sit unanswered for lack of staff capacity, as D04 reports when it raised the bar because submissions overwhelmed capacity.

**Step 2**

The endogeneity problem deserves its own step because it is the most likely place for a candidate to be caught. P06 states that better capabilities made empirical safety work more tractable, so that the main bottlenecks were no longer useful research paradigms but the availability of funding and talent. Read carefully, this is a claim about the funder's **own causal model**. The organization that reports the bottleneck is the organization that decided to triple spending and to shorten approval. The reallocation is strong evidence of a **belief** and weak independent evidence that money is genuinely binding rather than paradigm-limited.

The clean disambiguation: the bottleneck claim has two parts. Part one is **empirical** — money and talent are scarce relative to demand at this moment. Part two is **structural** — no further paradigm shift is needed to absorb more capital. A grantmaker should ask what observation would falsify part two, and the honest answer is that a subsequent cohort of grants failing to clear the bar, or returning to paradigm-limited, would do it.

*Inference: a funder that both diagnoses the bottleneck and supplies the remedy should be evaluated on the falsifiers it publishes, not on the diagnosis alone. D01's reasoning-transparency standard is the instrument here, because it forces the confidence and the kind of support to be declared.*

### Step 3

A second non-trivial move is that the same agenda is funded across several orders of magnitude, and the evidence required to establish a bottleneck is scale-dependent. If a $200k check is the largest check a grantee has ever received, the bottleneck claim can be settled by a reference and a pilot. If a $160m check is the largest AI safety grant ever made, the bottleneck claim must survive a claim about the grantee's capacity to absorb that capital without degrading.

Compute the stratification. The TAIS average grant is $157m divided by 200+, roughly $785k; the span from the smallest rapid grant to the Resolution grant is more than six orders of magnitude.

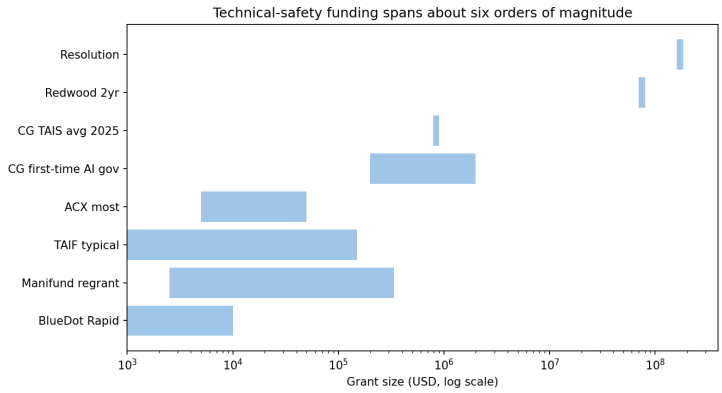


Figure 27: Graphical analysis for Step 3.

### Step 4

Now assign a bottleneck to each agenda, state the evidence offered by the sources, and flag endogeneity. The point of the table is that the evidence quality is not uniform: some bottlenecks are supported by an external, independently observed constraint, and others by the funder's own decision.

Table 54: Bottleneck claims per agenda and the endogeneity of the supporting evidence.

| **Agenda** | **Binding input claimed** | **Evidence offered** | **Endogeneity risk** |
|---|---|---|---|
| Evaluations | model access and independent funding | METR avoids lab funding to avoid conflicts; hiring is its central constraint (P07) | low |
| Control | people and money | Redwood scaling grant above $70m over two years (P01) | medium |
| Empirical safety | funding and talent | TAIS giving rose $46m to $157m across 200+ grants (P06) | high |
| Alignment and oversight | research paradigms (previously) | P06 says better capabilities made empirical safety more tractable | high |
| Security | execution capacity | P04 calls many infosecurity problems straightforward execution challenges | low |
| Interpretability | tooling and researchers | cross-lab SAE tooling collaboration (P11) | low |
| Red-teaming | coverage design | coverage, not resource, is the binding choice | low |
| Formal guarantees | theory and tooling | world-model, specification and verifier architecture; $30k seminar series (N04) | low |

## Sub-Question 3: How do funder priorities and the tractability dispute reshape the map?

Funder priorities are revealed by allocations and by the mechanisms built to receive proposals. The tractability dispute is partly a factual disagreement and partly an equivocation on what **tractable** means. Separate the two, and the map stops serving as a verdict and starts serving as a hedging instrument.

### Step 1

Map funder priorities by mechanism and scale, not by slogan. Two funders can both care about evaluations and still be non-substitutable if one makes $10k rapid grants and the other makes $2m renewal decisions.

Table 55: Revealed funder priorities across mechanisms and scales.

| Program | Mechanism | Scale | Agendas favored |
|---|---|---|---|
| CG TAIS | targeted RFP plus active grantmaking | $157m in 2025 across 200+ grants (P06) | empirical safety, evaluations, control, oversight |
| CG AI Governance RFP | rolling EOI across six subject areas | first-time $200k-$2m per year (D04) | technical governance, evaluations, weight security, standards |
| Redwood Research | direct large grant | above $70m over two years (P01) | control, alignment faking |
| Resolution | direct large grant | $160m, largest AI safety grant (P01) | rigorous alignment research |
| METR | recommended to individual donors | hiring-limited (P07) | evaluations, frontier safety policies |
| Guaranteed Safe AI seminars | small grant, seminar series | $30k (N04) | world-model, specification, verifier |

The convergence is that almost every program funds evaluations and measurement in some form. The divergence is in where the money lands once the measurement is done: control and alignment at the large end, field-building at the small end.

### Step 2

Disambiguate the word at the center of the dispute. **Tractable** is doing three jobs at once:

- Technically achievable: the research question can, in fact, be answered with available methods.
- Evaluable by the funder at acceptable cost: a grantmaker can tell within a grant cycle whether the work is on track.
- Absorptive of dollars at speed: the agenda can turn money into output without a paradigm bottleneck.

P06's claim that empirical safety became more tractable is a claim about senses one and two, with a practical consequence for sense three. The Guaranteed Safe AI approach in N04 scores high on the *claim* of sense one — fixed guarantees are more convincing than testing — but its current instrument is a $30k seminar series, so it is low on sense three. Evaluations are the reverse: high on senses two and three, and the open problem is whether the measurement has construct validity.

*Inference: a grantmaker should suspect the agenda rankings of the funder she is inspecting, because revealed priorities load on senses two and three, and epistemic merit lives in sense one. This is not a criticism of the funder; it is a reason to read the map twice with different weightings.*

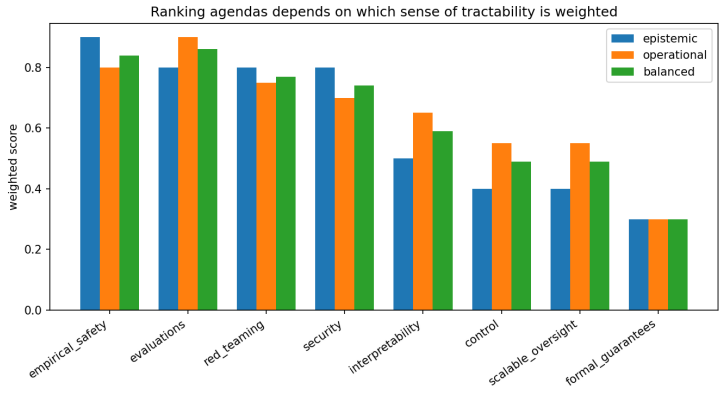


Figure 28: Graphical analysis for Step 2.

### Step 3

Run the three weightings and note that the top of the list is stable while the middle is not. Under the purely epistemic weighting, empirical safety, evaluations, red-teaming and security cluster at the top; under the operational weighting,

evaluations and empirical safety separate out and control rises relative to interpretability. The practical consequence: a funder's preference between control and interpretability is weak evidence about which is more important and strong evidence about which is easier to steward.

The second half of the dispute is not about the word at all. P08 quotes Luke Muehlhauser's 2020 statement of the underlying uncertainty: would tighter regulation reduce catastrophic risk or raise it by displacing development, and would accelerating AI help or hurt. That is a disagreement about the **causal structure**, not about tractability. A candidate who merges these two disputes into one will fail the interview's test of identifying the decisive consideration.

**Step 4**

The disagreement-tolerance principle is what makes the map usable. P08 says CG aims to support a wide range of experts, including those who disagree with each other and with CG, and it operationalizes this by funding both sides of a live dispute. The map is therefore a **hedging instrument**, not a verdict about the winning paradigm. A grantmaker's job is to keep the conditions under which the best paradigm can win, which in practice means protecting a set of programs whose success is negatively correlated.

The corollary for a candidate answer is a trap: it is tempting to name the best-supported agenda and fund only that. The source-supported discipline is the opposite — state the view, state the confidence and the kind of support, and state what would change your mind. That is D01 applied to a domain claim.

**Step 5**

Close with the open problems, because in a grant interview the open problems are the funding opportunities. Each row below converts a technical uncertainty into a decision about whether to fund, wait or refer.

Table 56: Open problems and their funding consequences.

| **Agenda** | **Open problem** | **Why it is decisive for funding** |
|---|---|---|
| Alignment | does a robust, learnable objective exist | determines whether alignment is research or engineering |
| Control | can control hold without solving alignment | determines whether control substitutes for or complements alignment |
| Interpretability | does internal understanding transfer to deployment safety | determines whether interpretability is safety work or science |

| **Agenda** | **Open problem** | **Why it is decisive for funding** |
|---|---|---|
| Scalable oversight | does oversight scale with capability or lag it | determines whether oversight is a race that can be won |
| Evaluations | construct validity of benchmarks | determines whether a clean eval result is decision-relevant |
| Red-teaming | coverage under adversarial search | determines whether absence of findings is evidence of safety |
| Security | threat model: insider versus outsider | determines which controls are worth funding |
| Empirical safety | whether the observed phenomenon proxies the catastrophic one | determines whether empirical findings generalize |

Caveat on sourcing. The agenda inventory, the funding figures, the control and alignment-faking work, the Guaranteed Safe AI architecture and the disagreement-tolerance principle are source-supported (P01, P04, P06, P07, P08, D04, N04). The classification by output type, the hypergraph framing, the complement relation between evaluations and control, the endogeneity ranking of bottleneck evidence, the six-order-of-magnitude stratification and the disambiguation of tractability into three senses are inferences constructed here for the interview. The exercise of naming a fundable bottleneck and its evidence is proposed, not attested in the corpus.

# Map the AI governance and policy landscape from research to enforcement.

## Question

Map the AI governance and policy landscape from research to enforcement. Connect research, policy recommendations, advocacy, lawmaking, regulation, standards, corporate compliance and enforcement. Provide example actors and outputs, and mark acknowledged gaps and uncertainties.

## Background

D04 sets the grantmaking scope. Eligible AI governance subject areas are technical AI governance, policy development, frontier company policy, international AI governance, law, and strategic analysis or threat modeling. AI governance work without likely major relevance to global catastrophic risks from transformative AI is ineligible, and work better fit for another RFP is also ineligible. This scope condition matters because a complete map of the field is not the same as a map of what CG will fund.

N06 supplies the most explicit supplied map. It lists 17 common AI governance activities, describes three pathways, Hard Governance, Industry-wide Self Governance, and Company Self Governance, and gives typical actors and outputs. It also states its own limits: the map is based mainly on experience in the EU, the US, the OECD, France, and the G7 Global Partnership on AI, and it is not accurate for every jurisdiction. Five macro-areas are explicitly left out: military and national security, supply chain and trade, multilateral soft governance, extralegal governance, and academic governance.

P08 adds CG's own policy and governance examples, including private governance initiatives, the if-then structure for frontier AI safety policies developed by Anthropic, METR, UK DSIT, OpenAI, Google DeepMind, and CEIP, Frontier Model Forum standards, national policy work, academic centers and think tanks, proposed laws such as the Preserving American Dominance in AI Act and California SB 53, and multilateral work. P04 shows that AI governance and policy grantmakers focus on active grantmaking, US policy, strategy, and grant write-ups.

*The non-trivial jump is not listing stages. It is seeing that the same policy idea can travel through public hard governance or private self-governance; that the binding force can come from law, procurement, reputation, market access, or internal authority; and that the most grantmaker-relevant gaps are often where the map is thin, jurisdiction-dependent, or missing an owner.*

## Sub-Question 1: Trace the AI Governance Pipeline and Locate Grantmaker-Relevant Leverage

Produce a map that connects research, policy recommendations, advocacy, law-making, regulation, standards, corporate compliance, and enforcement. Use example actors and outputs, then trace one policy idea through the pipeline. Finally, mark acknowledged gaps, jurisdiction dependence, and the points where a grantmaker would need decisive evidence.

### Step 1

**Step 1. Set the scope. The map must be filtered by catastrophic-risk relevance before it is used for grantmaking.**

D04 lists six eligible subject areas. They are not six pipeline stages. They are clusters of work that can appear at several stages of the pipeline.

- Technical AI governance: compute governance, model evaluations, technical safety and security standards, cybersecurity for model weights, and privacy-preserving transparency mechanisms.
- Policy development: government policy proposals developed in enough detail to be debated and implemented.
- Frontier company policy: voluntary or required practices such as model evaluations, scaling red lines and if-then commitments, incident reporting protocols, and third-party audits.
- International AI governance: paths to effective multilateral governance and coordination among key state actors.
- Law: liability, antitrust, legal tools for implementing high-level governance solutions, sound legal drafting, and legal aspects of policy proposals.
- Strategic analysis and threat modeling: threat models, takeoff speeds, timelines, reference class forecasting, and strategic clarity.

*Ineligible work includes AI governance without likely major relevance to global catastrophic risks from AI, and EOIs better fit for another active RFP. The practical test is whether the work could characterize or mitigate a catastrophic-risk pathway, not whether it is about AI policy in general.*

### Step 2

**Step 2. Inventory the 17 activities from N06. Read the table as an inventory, not as a causal model. The confidence column is N06's own stated confidence in its understanding of the activity group. Low-confidence nodes are where a grantmaker should be most careful about assuming the mechanism is understood.**

Table 57: Source: N06. Confidence is N06's stated confidence in its understanding of the activity group.

| **Activity** | **Example actors** | **Example outputs** | **Confidence** |
|---|---|---|---|
| Researching macro strategy | Academic centers, think tanks, funding agencies, government foresight units | Superintelligence, AI governance research agendas, forecasting reports | high |
| Applied research | GovAI, CSER, CLTC, OECD AI Policy Observatory, WEF | Malicious Use of AI report, CSET issue briefs, mappings of AI policy initiatives | high |
| Designing policy recommendations | CSET, Brookings, Wilson Center, lab policy teams | Internal policy briefs, policy recommendations, responses to public calls | high |
| Advocating and lobbying | Company government affairs teams, NGOs, lobbyists, universities | One-on-one meetings, public campaigns, open letters, stakeholder maps | high |
| Policymaking and lawmaking | Legislators, legislative staff, civil servants, policy advisors | Statutes, amendments, votes, funding directives | high |
| Regulations and executive guidance | Ministries, government agencies, executive offices | National AI strategies, guidance, executive directives | high |
| Standards-setting | ISO, ITU, CEN-CENELEC, ETSI, NIST, ANSI | Technical reports, standards, testing protocols, assessment methods | high |

| Activity | Example actors | Example outputs | Confidence |
|---|---|---|---|
| Industry coalition or network building | Trade associations, WEF, Partnership on AI, DigitalEurope | Roundtables, conferences, workshops, manifests | high |
| Industry norm building | CSR teams, PR teams, secretariats, institutional entrepreneurs | Commitments, principles, toolkits, new secretariats | low |
| Corporate compliance | Compliance teams, chief compliance officers, law firms | Standard operating procedures, training materials, compliance gap reports | low |
| Company internal coalition-building | Employees, investors, board directors, external advocates | Requests, pressure campaigns, legal and judicial advisory | low |
| Corporate governance | Board directors, C-suite, Responsible AI teams, AI safety teams | Mandates, internal policies, reporting lines, strategy changes | high |
| Public relations and communications | In-house communications teams, PR firms, marketing | Campaigns, crisis response toolkits, attitude surveys | high |
| Quality assurance and related | QA teams, product safety leads, internal control roles | Stress tests, technical documentation, quality assessments | low |
| AI development management | Line managers, product managers, engineers, HR and operations | Team scope, performance reviews, internal communications, culture | high |

| Activity | Example actors | Example outputs | Confidence |
|---|---|---|---|
| Enforcement and verification | Regulators, market surveillance authorities, auditors, conformity assessment bodies | Fines, warnings, investigations, audits, lawsuits | high |
| Evaluating policy | Research firms, think tanks, OMB, REFIT, ombuds institutions | Evaluation reports, policy datasets, budget recommendations | low |

**Step 3**

**Step 3. Trace one policy idea through Hard Governance. Take a concrete idea such as an if-then frontier safety policy or a compute governance mechanism and follow it through N06's Hard Governance pathway. The point is to see where the idea can lose force, where a new actor must be persuaded, and where grantmaker evidence would be decisive.**

1. Macro strategy research sets overall direction and priorities.
2. Applied research generates specific insight on the governance problem.
3. Policy recommendations turn insight into a jurisdiction-specific proposal.
4. Advocacy and lobbying transmit the recommendation to decision-makers.
5. Policymaking and lawmaking create a law or mandate.
6. Regulations and executive guidance specify the obligation.
7. Standards-setting translates the obligation into technical requirements.
8. Corporate compliance interprets the requirement into internal process.
9. Corporate governance adjusts strategy and internal authority.
10. Quality assurance adjusts testing and documentation.
11. AI development management changes developer behavior.
12. Enforcement and verification check the compliance trace, and policy evaluation feeds back into the next cycle.

P08 anchors the pathway with examples: the if-then structure for frontier AI safety policies developed by Anthropic, METR, UK DSIT, OpenAI, Google

DeepMind, and CEIP; Frontier Model Forum standards; national policy work; academic centers and think tanks; proposed laws such as the Preserving American Dominance in AI Act and California SB 53; and multilateral work.

*The caveat is that the path is not linear. Policy recommendations can be adopted directly by companies. Enforcement can be private certification or market surveillance rather than public regulation. Policy evaluation can be politicized and may not feed back cleanly. A grantmaker should therefore ask which stage the proposed work actually moves, and what evidence would show that the stage is the binding constraint.*

**Step 4**

**Step 4. Compare the three pathways. The same policy idea can travel through Hard Governance, Industry-wide Self Governance, or Company Self Governance. The pathway determines who must be persuaded, where the binding force comes from, and what would count as failure.**

Table 58: Source: N06's three macro-activities, with examples drawn from N06 and P08.

| Pathway | Initiator | Mechanism | Binding force | Example |
|---|---|---|---|---|
| Hard Governance | Governments and legislatures | Law, regulation, executive guidance, mandated standards | Legal coercion and enforcement | EU AI Act, national AI strategies, proposed US and California laws |
| Industry-wide Self Governance | Industry coalition or network | Norms, standards, industry commitments | Reputation, procurement, market access, peer pressure | Frontier Model Forum standards, Partnership on AI, WEF initiatives |
| Company Self Governance | Internal coalition of employees, investors, or leaders | Corporate governance, internal policy, compliance, QA | Employment, investment, internal authority | Responsible AI teams, if-then safety policies, internal reporting |

*The grantmaker implication is that a proposal may be aimed at the wrong path-*

*way. Funding research may not help if the bottleneck is coalition-building. Funding a standard may not bind if procurement does not adopt it. Funding a company policy may fail if the internal coalition lacks authority. P08's willingness to fund a wide range of actors, including those who disagree, is one response to this uncertainty.*

**Step 5**

**Step 5. Map D04's six subject areas onto the pipeline and example actors. D04's subject areas are not stages. They are clusters of work that can appear at several stages, and the grantmaker question is whether the work moves a binding constraint or only describes the landscape.**

Table 59: Sources: D04 eligible subject areas; P08 and N06 for examples.

| **D04 subject area** | **Typical pipeline stages** | **Example actors or outputs** | **Grantmaker-relevant question** |
|---|---|---|---|
| Technical AI governance | Applied research, standards, compliance, QA, enforcement | Compute governance, model evaluations, technical safety and security standards, weight security, privacy-preserving transparency | Does the mechanism change the feasibility or efficacy of governance? |
| Policy development | Recommendations, lawmaking, regulation | Policy briefs, draft proposals, implementation prerequisites | Is the proposal specific enough to be debated and implemented? |
| Frontier company policy | Industry norms, corporate governance, compliance | Evaluations, red lines, if-then commitments, incident reporting, third-party audits | Would voluntary adoption bind, or would it be gamed? |

| D04 subject area | Typical pipeline stages | Example actors or outputs | Grantmaker-relevant question |
|---|---|---|---|
| International AI governance | Macro strategy, coalitions, multilateral work | International institutions, coordination mechanisms, state-actor pathways | Is there a plausible path through state actors? |
| Law | Recommendations, lawmaking, enforcement | Liability, antitrust, legal tools, drafting of impactful policies | Does legal drafting create a real lever? |
| Strategic analysis and threat modeling | Macro strategy, applied research | Threat models, takeoff speeds, reference class forecasting | Does this change priorities, or only describe the landscape? |

P08 is useful here because it shows CG treating these clusters as a portfolio. It funds private governance, frontier safety policy structure, standards work, national policy, academic centers, proposed legislation, and multilateral research. It also funds people who disagree with each other and with CG, which makes the portfolio an uncertainty hedge as well as a set of bets.

**Step 6**

**Step 6. Mark acknowledged gaps, jurisdiction dependence, and grantmaker implications. The map is useful only if its missing parts and low-confidence nodes are visible.**

N06 explicitly omits five macro-areas. These are not peripheral to catastrophic risk, which makes them candidate areas for further research or grantmaking attention.

- Military and national security governance: defense actors, vetting, training, secrecy, intelligence, laws of war, and cyberweaponry.
- Supply chain and trade governance: contractual agreements, export controls, investors arbitration, standard contractual clauses, and trade authorities.
- Multilateral soft governance: UNESCO AI Principles, Asilomar Principles, G20 declarations, and OECD AI Principles.
- Extralegal governance: IP theft, cybercriminality, trafficking of cyberweapons, commercial espionage, and self-governance of criminal or hacker groups.

- Academic and R&D governance: research funding, program setting, appointments of university or AI department leaders.

*N06's own epistemic status limits the map. It is based mainly on EU, US, OECD, France, and G7 experience and interactions with AI developers in the US, UK, and multinationals. It is not accurate for every jurisdiction, and N06 gives the example that advocacy and lobbying differ in China. Any grantmaker should therefore treat the map as partial and jurisdiction-dependent rather than universal.*

Grantmaker implications are practical. D04 makes catastrophic-risk relevance the scope condition and treats some AI governance work as ineligible. P04 shows grantmaker time split across meetings, research, and grant write-ups, with a US policy focus for one of the profiled grantmakers. P08 shows that CG funds a wide range of actors and accepts disagreement. The map should therefore be used to locate decisive evidence, route proposals to the right pathway, and notice missing owners, not to demand a complete description of the field.

Proposed exercise. Choose one policy idea and trace it through the pipeline. At each stage, name the actor, the output, the binding force, the evidence that would change a grant decision, and the first gap where no owner exists. Then repeat the trace through Company Self Governance and Industry-wide Self Governance. Source-supported material supplies the activities, actors, and examples. The staged trace is an exercise, and the claim that gaps are leverage is an inference to test rather than a finding from the sources.

# Map key AI safety and security organizations, funders and field-building infrastructure.

## Question

Produce an ecosystem map of major AI safety and security organizations, funders, field-building programs, and evaluators, with roles, relationships, and evidence gaps. The map must be usable in a grantmaker interview: it should let you say who does what, who funds whom, where the handoff regions are, and where the map is uncertain or contested.

*The non-trivial requirement is not to list organizations but to model the functional pipeline: how opportunities, talent, money, evaluations, and policy outputs move through the field. A categorical list (lab, funder, nonprofit) collapses decision-relevant differences and hides multiplex roles.*

## Background

Sources are snapshots. P01 names Redwood Research, Epoch AI, BlueDot Impact, 80,000 Hours, Project Tailwind, and Resolution, and describes CG's scaling and grantmaker roles. P06 gives MATS context and early-bet compounding. P07 suggests METR and notes nine major AI companies have frontier safety policies, with METR credited for Google DeepMind, Anthropic, Magic, G42, and Amazon, and METR avoiding lab funding to prevent conflicts. P08 names OpenAI, Google DeepMind, Anthropic, METR, UK DSIT, Carnegie Endowment, and the Frontier Model Forum. P09 describes BlueDot Context Week. P10, P11, P13, and P14 describe EA Funds, Manifund, and BlueDot rapid grants. N01-N06 describe events, Modeling Cooperation, AI Safety Poland, GSAI seminars, MIT AI Risk Initiative, and governance mapping. S03 describes Incubator Week outputs and follow-on pre-seeds from Coefficient Giving, Longview, and BlueDot.

*Separate source-supported claims from inferences. The sources support many co-funding and role relations, but they do not provide a complete census, current staffing, or a full grant graph. Organizational roles, funding, and staffing change quickly. Any map built here is a dated, partial model, not the territory.*

## Sub-Question 1: Mapping axes: from categorical labels to functional roles

Before listing organizations, define the dimensions along which they differ for a grantmaker. A categorical list (lab, funder, nonprofit) collapses decision-relevant differences. What functional roles should the map encode, and how do you handle entities that occupy several roles at once?

**Step 1**

*The non-trivial jump: an ecosystem map is not a directory. It is a model of the grantmaking pipeline and the field's capacity. The unit of analysis is the role, not the legal form. An organization can be simultaneously a research producer, an evaluator, a grantee, a regranter, and a standard-setter.*

- Function in the decision pipeline: who originates opportunities, who evaluates, who recommends, who decides, who pays, who executes.
- Stage of the talent and idea lifecycle: pre-idea, idea formation, founder formation, early project, scaling, consolidation.
- Funding instrument and ticket band: rapid microgrant, first institutional grant, scale-up.
- Independence and conflict posture: whether the entity accepts money from labs it evaluates or regulates.
- Jurisdiction and policy lever: domestic law, executive guidance, standards body, international coordination.
- Community and trust position: insider vs outsider, EA-affiliated vs adjacent, open vs secretive.

*The map should be multiplex. A single entity appears multiple times with different roles. That is not a failure of classification; it is the phenomenon to be captured.*

| **Axis** | **Why it matters for a grantmaker** | **Example entities (source-supported)** |
|---|---|---|
| Function in decision pipeline | Tells you where an entity sits in the causal path from idea to grant to outcome | CG (funder/incubator), METR (evaluator), Manifund (regranter), BlueDot (funder and field-builder) |
| Stage of lifecycle | Different stages have different bottlenecks and funding instruments | Project Tailwind (founder formation), Incubator Week (early project), MATS (talent pipeline), Resolution (scaling) |
| Instrument and ticket band | Determines who can fund what and where seams appear | BlueDot up to $10k (D03), Transformative AI Fund below $150k typical (P10), CG first-time $200k-$2m/yr (D04) |

| Axis | Why it matters for a grantmaker | Example entities (source-supported) |
|---|---|---|
| Independence posture | Affects credibility of evaluations and policy recommendations | METR does not take lab funding (P07) |
| Jurisdiction and policy lever | Policy organizations are not interchangeable across jurisdictions | UK DSIT (P08), EU and US actors in N06, MIT AI Risk Initiative (N05) |
| Community and trust position | Determines dealflow and ability to recruit | P09 on community norms, P12 on trust as bottleneck |

**Step 2**

*Define roles by what the entity does to the grant decision or field capacity. Use verbs, not nouns.*

- Funder/decider: holds money and makes final grant decisions (CG, Good Ventures, EA Funds Transformative AI Fund, Manifund regranting pool, BlueDot Rapid Grants, SFF, Longview).
- Recommender/intermediary: evaluates and recommends but does not hold final money (GiveWell for global health, Manifund regrantors, ACX Grants evaluators).
- Evaluator: produces assessments used by funders or policymakers (METR, Apollo Research, MIT AI Risk Initiative, GiveWell).
- Research producer: creates technical or strategic knowledge (Redwood Research, Resolution, Epoch AI, Apollo Research, Modeling Cooperation, academic centers).
- Field-building/talent infrastructure: recruits, trains, places, or connects people (BlueDot Impact, 80,000 Hours, MATS, Project Tailwind, Incubator Week, AI Safety Poland, career advising programs).
- Policy and governance organization: produces policy analysis, recommendations, or implementation (GovAI, CSET, Carnegie Endowment, UK DSIT, Frontier Model Forum, MIT AI Risk Initiative).
- Standard-setter or industry self-governance body: translates requirements into technical specifications or voluntary commitments (Frontier Model Forum, METR's frontier safety policy work, industry coalitions).
- Fiscal sponsor or infrastructure provider: handles legal, financial, or operational plumbing (EA Poland for AI Safety Poland, Manifund for payouts, Convergence Analysis for Modeling Cooperation).

*Some roles are hierarchical; others are orthogonal. A funder can also be a field-builder (CG's Project Tailwind). A research producer can also be an evaluator (METR). A field-building program can become a funder (BlueDot Rapid Grants). The map should record role overlaps explicitly.*

**Step 3**

Apply the taxonomy to a few source-supported entities to show multiplexity.

| **Entity** | **Roles** | **Source-supported evidence** | **Inference or gap** |
|---|---|---|---|
| Coefficient Giving | Funder, recommender, incubator, standard-reinforcer | P01, D04, S09 | Inference: standard-reinforcement from P01 footnote on best practices |
| METR | Evaluator, policy shaper, research producer | P07, P08 | Inference: research producer from its evaluations and policy work |
| BlueDot Impact | Field-builder, funder (Rapid Grants), community | P09, P13, P14, D03 | Source-supported |
| Manifund | Funder (regranting), infrastructure provider, fiscal sponsor | P11, P12 | Source-supported |
| Apollo Research | Evaluator, research producer | P11 | Source-supported |
| Epoch AI | Research producer, data source | P01, P06 | Source-supported |
| Resolution | Research producer, grantee | P01, P11 | Source-supported |
| AI Safety Poland | Field-builder, grantee | N03 | Source-supported |
| Modeling Cooperation | Research producer, fiscal sponsor client | N02 | Source-supported |

*Evidence gaps: the sources do not give a complete list, current staffing, or formal governance relations. Many edges are inferred from co-funding, merging, or shared personnel. The map is a snapshot; update triggers are new RFPs, fund launches, mergers, or large grants.*

## Sub-Question 2: Funders and funding instruments: ticket bands, decision speed, and trust gates

Map the funders and instruments that move money into AI safety and security. What are the source-supported ticket bands, decision speeds, openness, and trust models? Where do the seams and handoff regions appear?

### Step 1

Start with the money. The following table encodes source-supported attributes. Do not read it as a complete census.

| **Funder** | **Role** | **Source-supported ticket band** | **Decision speed / openness** | **Trust model / notes** |
|---|---|---|---|---|
| Coefficient Giving | Funder, recommender, incubator | First-time grants typically \$200k-\$2m/yr over 1-2 years (D04) | Faster decisions, one active sign-off besides primary investigator (P01) | Trust inherited via GiveWell, now institutional (P12); funds people who disagree (P08) |
| Good Ventures | Largest funding partner, partner organization | Gives via CG recommendations (D04) | Not separately described | Partner to CG |
| EA Funds Transformative AI Fund | Early-stage funder | Typically below \$150k; rare above \$300k; no >~\$1m foreseeable (P10) | Open applications, quarterly reports, seed ~\$1m | New full-time team; bets on field absorption (P10) |

| Funder | Role | Source-supported ticket band | Decision speed / openness | Trust model / notes |
|---|---|---|---|---|
| Manifund regranting | Regrantor pool | 2025 $2m across 13 regrantors; top hits $2.5k-$335k (P11) | Part-time regrantors; donors cover 5% op cost | Trust via regrantor track records and transparency (P12) |
| BlueDot Rapid Grants | Funder (rapid grants) | Up to $10k; typical $50-$1,500 (D03, P13, P14) | Average reply within a day; 9 in 10 within a week; upfront by default | Trust via community eligibility; work already in motion (D03) |
| Survival and Flourishing Fund | Funder, matching facility | Matched grant for Modeling Cooperation (N02) | Not fully described | Committed grant unlocks match (N02) |
| Longview Philanthropy | Co-funder | Co-funded Incubator Week pre-seeds (S03) | Not described | Appears as co-funder (S03) |
| ACX Grants | Microgrants program | Most $5k-$50k; rare up to $100k; ~$1m round (P16) | Annual round; evaluator-driven | Trust banked in Scott Alexander (P12) |
| LTFF (wound down) | Historical early-stage funder | 820+ grants, just under $35M since 2017; mean ~$42k (P10) | Closed; transitioned to Transformative AI Fund | Legacy niche: early-stage, small, speculative (P10) |

**Step 2**

Ticket bands are not just sizes; they create seams and handoff regions. D04 says projects below $200k/yr should ideally be funded by other funders. P10 says the Transformative AI Fund typically funds below $150k and rarely above $300k. That leaves a visible handoff region around $150k-$200k, and a scale-up region above $2m that CG may enter after evidence.

*The deeper non-trivial point is that trust is the gate, not money. P12 argues trust will be the bottleneck for deploying the coming funding wave. Different instruments convert trust into different mechanisms. BlueDot converts trust into an eligibility rule (community membership). Manifund converts trust into regrantor track records. CG converts trust into institutional reputation and a small army of specialised grantmakers. The map should therefore encode, for each funder, the trust-conversion mechanism, because that determines who can access the money.*

**Step 3**

The table below makes the trust-conversion mechanisms explicit.

| Instrument | Ticket band | Speed | Openness | Trust conversion mechanism | Source |
|---|---|---|---|---|---|
| BlueDot Rapid Grants | $50-$10k | 1 day average | Open to community | Eligibility rule | D03 |
| Transformative AI Fund | ~$150k typical | Quarterly reports | Open applications | New team plus field-absorption thesis | P10 |
| Manifund regranting | $2.5k-$335k | Part-time; variable | Donor-directed | Regrantor track record | P11, P12 |
| CG first-time | $200k-$2m/yr | Faster decisions | Active grantmaking plus RFPs | Institutional trust and grantmaker expertise | D04, P01, P12 |
| ACX Grants | Most $5k-$50k | Annual | Open | Trust banked in Scott Alexander | P16, P12 |

**Step 4**

Visualize the ticket bands to make the seams obvious. The code computes the handoff region between the Transformative AI Fund typical maximum and CG first-time minimum, and the graphical analysis plots all bands on a log scale.

```
taif_typical_max = 150000
cg_first_time_min = 200000
```

```
handoff_gap = cg_first_time_min - taif_typical_max
handoff_ratio = cg_first_time_min / taif_typical_max
output_state['handoff_gap_usd'] = handoff_gap
output_state['handoff_ratio'] = handoff_ratio
```

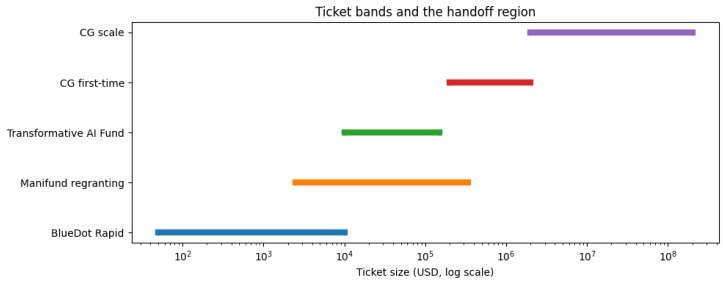


Figure 29: Graphical analysis for Step 4.

### Step 5

Evidence gaps for the funder map: actual annual giving to AI safety by each funder; overlap and double counting; additionality; whether ticket bands are binding constraints or descriptions; the handoff region may be filled by undisclosed funders; decision quality and error rates are unmeasured.

- P12 argues trust is the bottleneck; P10 argues the field lacks organizations; P01 argues staff time is the main constraint. These are disagreements about what binds first.
- S10 lists philanthropic funding pitfalls: fund based on cost not value delivered, game-of-chicken between funders, and unclear credit allocation.

## Sub-Question 3: Field-building and talent infrastructure: the reflexive layer

Map the field-building programs, talent pipelines, and community infrastructure. How do these programs relate to funders, and why is the field-building layer reflexive for a grantmaker?

### Step 1

Field-building programs are both grant targets and dealflow sources. This table maps source-supported programs.

| Program | Type | Role | Relationship to funders | Source |
|---|---|---|---|---|
| MATS | Fellowship | Talent pipeline for AI safety | CG early grant $250k in 2019; now ~100 scholars/year, 4-7% admit, >80% alumni in AI safety | P06 |
| BlueDot Impact | Courses, Context Week, Rapid Grants | Field-building, community, rapid funding | BlueDot itself funds rapid grants | P09, P13, P14, D03 |
| 80,000 Hours | Career advice and job board | Talent pipeline and information | Named by P01 as resource | P01 |
| Project Tailwind | Founder call | Incubator for new AI safety initiatives | CG initiative; funding at pre-seed, seed, scale levels | P01, S09, D04 |
| Incubator Week | Incubator | Early-stage startup formation | Outputs raised pre-seeds from CG, Longview, BlueDot | S03, S05 |
| AI Safety Poland | Local field-building | Community, events, career consulting | BlueDot Rapid Grant $4800 | N03 |
| GSAI Seminars | Seminar series and community hub | Research coordination | Manifund grant $30k | N04 |
| MIT AI Risk Initiative | Research and evidence synthesis | Decision support | Funded by Open Philanthropy, NSF, Microsoft, etc. | N05 |

| Program | Type | Role | Relationship to funders | Source |
|---|---|---|---|---|
| Effective Careers RFP grantees | Career advising and fellowships | Talent transitions | CG Effective Giving & Careers Fund | P02 |

### Step 2

*The field-building layer is reflexive: funders create the field they later fund. CG's early bets on MATS and YIMBY are cited as compounding (P06). BlueDot's Rapid Grants fund projects that then apply to CG or Manifund (P13, N03). Incubator Week outputs raise pre-seeds from CG, Longview, and BlueDot (S03). This means the map's edges include talent flows and dealflow flows, not just money flows.*

*A second non-trivial point: field-building organizations often become grantmakers themselves, blurring the funder/grantee distinction. BlueDot runs courses and funds rapid grants. Incubator Week is run by someone who is also a grantmaker (S03: 'accelerated AI safety grantmakers (me included)'). Manifund is a funding platform and a fiscal sponsor. The map should record these dual roles.*

### Step 3

The network diagram makes the reflexivity visible. The code draws source-supported funding and talent-flow edges among the entities in this map.

```
import networkx as nx
import matplotlib.pyplot as plt

G = nx.DiGraph()
edges = [
    ('Coefficient Giving', 'MATS'),
    ('Coefficient Giving', 'Redwood Research'),
    ('Coefficient Giving', 'Epoch AI'),
    ('Coefficient Giving', 'Resolution'),
    ('Coefficient Giving', 'Project Tailwind'),
    ('BlueDot Impact', 'AI Safety Poland'),
    ('Manifund regranting', 'Timaeus'),
    ('Manifund regranting', 'Apollo Research'),
    ('Manifund regranting', 'Neuronpedia'),
    ('Manifund regranting', 'ChinaTalk'),
    ('SFF', 'Modeling Cooperation'),
    ('Longview Philanthropy', 'Incubator Week outputs'),
    ('BlueDot Impact', 'Incubator Week outputs'),
    ('Coefficient Giving', 'Incubator Week outputs'),
]
G.add_edges_from(edges)
pos = nx.spring_layout(G, seed=42)
```

```
nx.draw(G, pos, with_labels=True, node_size=2000,
    node_color='lightblue', font_size=8, arrows=True)
plt.tight_layout()
plt.savefig(_generated_image_path_)
```

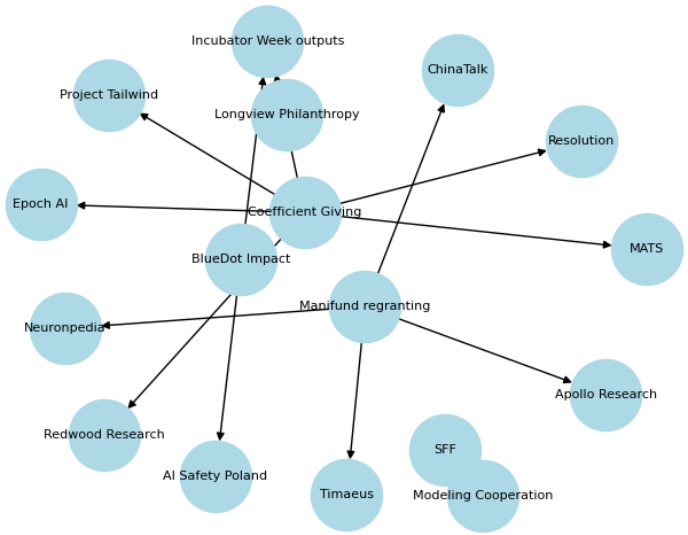


Figure 30: Graphical analysis for Step 3.

### Step 4

Evidence gaps: actual talent outcomes and counterfactual career transitions; whether field-building is saturated; duplication among career advising programs; the degree to which field-building programs create new opportunities versus redirect existing talent; the causal effect of community and events.

## Sub-Question 4: Research, evaluation, and policy organizations: producers, evaluators, and standard-setters

Map the organizations that produce research, evaluate systems or policies, and set standards. What functional distinctions matter, and where is the map jurisdiction-dependent or contested?

### Step 1

This table separates three functions that are often conflated: producing research, evaluating systems or policies, and setting standards.

| Organization | Primary function | Key output/role | Relationships/notes | Source |
|---|---|---|---|---|
| Redwood Research | Technical research lab | Alignment-faking, AI control | CG grant >$70m over two years | P01 |
| Resolution | Technical research lab | Rigorous alignment research | $160m CG grant; Timaeus merged into it | P01, P11 |
| METR | Evaluator and policy shaper | Frontier AI safety policies; capability evaluations | Nine major AI companies have policies; credited with Google DeepMind, Anthropic, Magic, G42, Amazon; no lab funding | P07 |
| Apollo Research | Evaluator | Empirical evaluations for scheming | Early Manifund grant $335k; grew to ~40 people | P11 |
| Epoch AI | Research producer | Public source of truth on AI trajectory | $68m renewal over 3 years | P01 |
| OpenAI, Google DeepMind, Anthropic | Frontier labs | Develop frontier models; acknowledge control problems; have policy teams | Named in P08 | P08 |
| UK DSIT | Government | If-then policy structure | Named in P08 | P08 |
| Carnegie Endowment | Think tank | If-then structure | Named in P08 | P08 |

| Organization | Primary function | Key output/role | Relationships/notes | Source |
|---|---|---|---|---|
| Frontier Model Forum | Industry standards body | Safety standards development | Named in P08 | P08 |
| GovAI, CSET | Policy research | Governance research and recommendations | Named in N06 | N06 |
| MIT AI Risk Initiative | Research and evidence synthesis | Risk repository, incident tracker, mitigations database | Funded by Open Philanthropy, NSF, Microsoft, etc. | N05 |
| Modeling Cooperation | Research producer | Simulation software and competition dynamics research | SFF matched grant; fiscally sponsored by Convergence Analysis | N02 |

**Step 2**

*Distinguish research production from evaluation from standard-setting. METR does all three but has a structural independence rule: it does not take lab funding (P07). That rule is a non-trivial design feature: it converts independence into a credibility asset and a funding constraint. Standard-setting bodies like the Frontier Model Forum are not evaluators; they translate requirements into technical specifications (P08).*

*Policy maps are jurisdiction-dependent and partial. N06 explicitly excludes military and national security governance, supply chain and trade, multilateral soft governance, extralegal governance, and academic governance. It also warns that explanations will not be accurate for some jurisdictions, e.g., advocacy and lobbying differ in China. Any policy map built for a grantmaker should state its jurisdiction and exclusions.*

*CG explicitly funds people who disagree with each other and with CG (P08). For example, P08 describes funding both a paper suggesting language models could help develop pathogens and a RAND paper finding current models did not substantially increase biological attack risk. The map should show disagreement, not merge organizations into one voice.*

### Step 3

A simple count by function makes the distribution visible. The chart is illustrative and based on the table above, not a census.

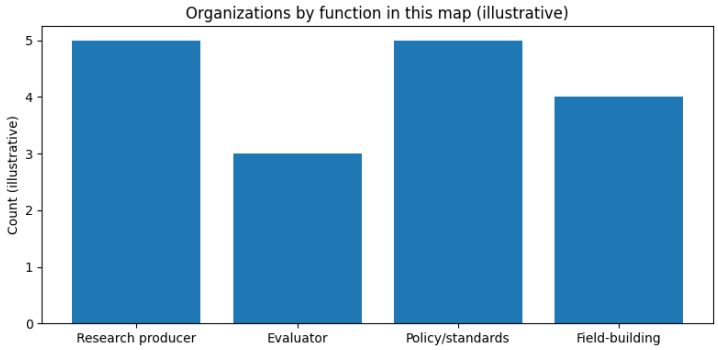


Figure 31: Graphical analysis for Step 3.

### Step 4

Evidence gaps: actual influence of policy organizations; whether METR's no-lab-funding policy is sustainable; overlap between policy orgs; coverage of non-US/EU jurisdictions; the degree to which evaluations change lab behavior; the pipeline from research to policy adoption.

## Sub-Question 5: Evidence gaps, disputes, and maintenance rules for the map

What are the explicit evidence gaps in this map? Who disputes the map or its organizing assumptions? What rules keep the map usable as sources age?

### Step 1

The map is only as good as its source dates. List the gaps explicitly.

- Funding volumes and staffing are snapshots; sources are dated.
- Relationships are often inferred from co-funding or merging, not governance.
- Counterfactual additionality is unknown: would the grant have happened anyway?
- Talent flows and career outcomes are not measured in a comparable way.
- Non-US/EU coverage is thin.

- Double counting through intermediaries is possible.
- The map may miss informal networks and unannounced projects.
- *The map embeds a theory of what matters; it is not neutral.*

**Step 2**

*A map that hides disagreement is less useful. Record who disputes what.*

| **Dispute** | **Who holds it** | **Source** |
|---|---|---|
| Open weights and transparency | Austin Chen argues open weights seem good, favors total research transparency, open data, full open source; OpenAI seems correct about being open; Anthropic seems good; not afraid of China; politics bad for EA soul; secrecy bad | S11 |
| Governance map completeness | N06 authors state map is partial, jurisdiction-dependent, excludes military, supply chain, multilateral soft governance, extralegal, academic | N06 |
| Biorisk and AI | P08 funded both a paper finding language models could help develop pathogens and a RAND paper finding current models did not substantially increase biological attack risk | P08 |
| Binding constraint | P12 argues trust is the bottleneck; P10 argues the field lacks organizations; P01 argues staff time is the main constraint | P12, P10, P01 |

| Dispute | Who holds it | Source |
|---|---|---|
| Field-building norms | P09 describes Context Week as questioning assumptions and admitting uncertainty; notes community norms matter | P09 |
| Philanthropic funding pitfalls | S10 lists fund based on cost not value delivered, game-of-chicken between funders, unclear credit allocation | S10 |

**Step 3**

*Treat the map as a living document with explicit update triggers and confidence labels.*

- New RFP or fund launch.
- New large grant or merger.
- Change in lab funding policy or conflict-of-interest rule.
- New jurisdiction or policy instrument.
- New evaluation result or capability demonstration.
- Organizational wind-down or leadership change.

For each entry, record: name, roles, source, source date, confidence, disputes, and what would change the entry. Use D01 reasoning transparency: state confidence and kind of support. Use the two denominators from the prior phase: expected value per dollar and expected value per grantmaker attention-hour. The map should help you decide where to spend attention, not just describe the field.

# Analyze AI safety community norms, context and strategic disagreements.

## Question

Produce a norms-and-disagreements brief for the AI safety, security and governance field: what the community means by **context**, what its **norms** are and what work those norms do, and where reasonable actors disagree, with cruxes preserved rather than resolved.

Success criteria. Every disagreement entry must name the **earliest point of disagreement**, state each side in its strongest recognizable form, name the evidence that would move it, and say what is *not* in dispute. No organization and no person may be merged into a single voice. Source-supported claims, inferences and proposed exercises must be labelled separately.

Constraint. The evidence base is a set of documents written for external audiences, not ethnography and not interviews. Any claim of the form *the community thinks* is an artifact of summarization unless a document is named that actually says it. P08 states that Coefficient Giving funds people who disagree with each other and with Coefficient Giving, so a unified institutional voice is often a compression error rather than a finding.

## Background

The reference role asks for familiarity with catastrophic-risk communities, the ability to compare opposing arguments, to reason quantitatively and probabilistically, to identify decisive considerations, and to communicate concisely and with transparent reasoning. Three of those five requirements are norms-of-argument requirements rather than domain facts. This brief treats norms and disagreement as first-class domain content, because the evaluation craft from the previous phase (crux identification, confidence and support kinds, proportionate diligence) is applied here to a community whose central coordination problem is **trust under growth**.

P12 supplies the mechanism. It describes the AI safety and effective-altruism funding ecosystem as starting with high mutual trust because it was small, weird, demanding and socially interwoven, and predicts that this condition will not scale. It calls trust the bottleneck to deploying the coming funding wave and identifies a principal-agent problem between the party holding money and the party holding taste. Those two observations do work here: if trust substitutes for verification, then **norms are a trust technology**, and a prediction follows about which norms survive contact with scale.

P08 and D04 make disagreement a design feature rather than a failure. P08 says the funder deliberately supports grantees who disagree about the nature of AI risk and about how to mitigate it, and it gives a worked example: two funded papers that point in opposite directions on whether language models materially

increase biological attack risk. D04 says strong applicants will often hold views that differ significantly from most staff who work on AI. A brief that resolves these disagreements into a consensus would misrepresent the institution it is meant to describe.

The most explicit list of community disagreements in the corpus is S11, an informal sketch of proposed essays titled *my disagreements with AI safety hivemind*. It supplies positions, not arguments, and its value is as a map of live fault lines: open weights seem good, transparency should be maximal, OpenAI was right about being open, Anthropic seems good, a pause or pace-by-default world may already be here, timelines may be medium or broad, general fearmongering seems bad, politics is bad for the soul of effective altruism, the fear of China is not understood by the author, and pluralism and tolerance are goods. Each of these is a candidate disagreement entry; none of them is evidence about the world.

Finally, the brief must be usable in a live conversation. The overall task forbids inventing credentials, work history or community membership. Everything below is therefore framed as reasoning about documented positions, and the interview-facing material is rehearsal of structure, not rehearsal of a biography.

## Sub-Question 1: What context actually is, and why it cannot simply be taught

Give the field's own definition of context, decompose it into components, classify which components are buildable by study and which are not, and explain what follows for how context is demonstrated rather than asserted.

### Step 1

P09 gives the only explicit definition of context in the corpus: it is **the knowledge and judgment that help someone understand the field and decide how to contribute**. That sentence contains two objects, not one: a knowledge object and a judgment object. P09 then decomposes the intuition into five items.

- Situational awareness of AI progress.
- Familiarity with different strategic agendas.
- The ability to evaluate and develop strategies independently.
- Knowledge of the organizations in the field and where one's own skills might fit.
- An understanding of the norms of AI safety and adjacent communities.

The first non-trivial move is to notice that the list is **heterogeneous in kind, not merely in topic**. Items one and four are retrieval: they can be looked up, and their absence is visible. Item two is a comparative map: it requires holding

several agendas at once without collapsing them. Item three is a skill: it is only observable in the performance of a strategy. Item five is tacit: it is acquired by immersion, and its absence is usually invisible to the person who lacks it. P09 concedes the split in its own words, that some components can be developed through focused study and practice while others require weeks of turning ideas over in a dense intellectual environment. This is the source grading its own list, and it matters because the components cannot be taught or tested by the same instrument.

Table 60: P09's five components restated as claims that could be observed. The fifth column is our inference from the other columns, not a source claim.

| Component | What it is | How it is acquired | Observable evidence of possession | Signature of absence |
|---|---|---|---|---|
| 1. Situational awareness of AI progress | Tracking capability trends and their implications | Focused study of public measurement | Cites dated measurements and their direction, e.g. agent task time horizons, capability indices, data-centre capital expenditure | Recites headlines without magnitude, date or direction |
| 2. Familiarity with strategic agendas | The ability to state several agendas in their strongest form | Reading plus conversation | Names what each agenda requires and what would falsify it | Collapses all agendas into one preferred view |
| 3. Independent evaluation and strategy development | Producing a strategy together with its failure conditions | Practice, feedback, observed results | Delivers a strategy with a crux, an update rule and a scope limit | Endorses strategies with no stated conditions under which they fail |

| Component | What it is | How it is acquired | Observable evidence of possession | Signature of absence |
|---|---|---|---|---|
| 4. Organizational knowledge and fit | Knowing who does what and where one's own skills fit | Direct contact and referral | Names organizations, their function and a specific unfilled gap | Names only the most visible organizations |
| 5. Understanding of community norms | Knowing what makes an argument persuasive here | Immersion in the community | Hedges confidence, cites kinds of support, states what is not in dispute | Overstates, merges voices, or imports norms from an adjacent field |

The second move is to notice that P09's diagnosis is partly **demand-side**. Its illustration of what a high-context person can do is not the retrieval of a fact but the anticipation of a need, described as setting up a rogue-agent hotline in May 2026 ten weeks before internal agents broke onto the open internet. Context therefore includes knowing what the field will need before anyone asks for it. That is the same faculty that S09 appeals to when it says there are critical basic problems that no one owns, and that some sources estimate the range of unmet demand rather than the state of supply.

The source also leaves its own weighting problem open, saying it wants to work out which components matter most for particular roles and what the field most needs. Any claim that one component dominates is therefore an inference. Make that inference explicitly, and mark it as one.

```
import math

# P09's five context components, annotated with properties that matter for a
# norms brief: how each is acquired, how it is observed, and how its absence
# shows up. teachability: 1 = study alone can build it, 5 = immersion only.
# decision_relevance: 5 = highest for grantmaking judgement.
CONTEXT_COMPONENTS = [
    {"id": "situational_awareness", "label": "Situational awareness of AI progress",
     "acquisition": "focused study", "teachability": 1, "decision_relevance": 3,
     "observable": "cites dated measurements and their direction",
     "failure": "recites headlines without magnitude or date"},
```

```
    {"id": "strategic_agendas", "label": "Familiarity with different strategic agendas",
     "acquisition": "reading plus conversation", "teachability": 3, "decision_relevance": 5,
     "observable": "states several agendas in their strongest form",
     "failure": "collapses all agendas into one preferred view"},
    {"id": "independent_evaluation", "label": "Ability to evaluate and develop strategies independently",
     "acquisition": "practice, feedback, results", "teachability": 4, "decision_relevance": 5,
     "observable": "produces a strategy with a crux and an update rule",
     "failure": "endorses strategies with no stated failure conditions"},
    {"id": "organizational_knowledge", "label": "Knowledge of organizations and where one's skills fit",
     "acquisition": "direct contact and referral", "teachability": 2, "decision_relevance": 3,
     "observable": "names organizations, their function and a specific gap",
     "failure": "names only the most visible organizations"},
    {"id": "community_norms", "label": "Understanding of norms of the community",
     "acquisition": "immersion", "teachability": 5, "decision_relevance": 4,
     "observable": "hedges confidence, cites support kinds, states disagreements",
     "failure": "overstates, merges voices, imports foreign norms"},
]

def context_profile(components):
    """Rank components by how invisible an outsider's own gap is likely to be.

    looseness = how far the component is from being buildable by study alone.
    The product is ordinal. It is not a measurement and it must not be read
    as one; it only orders which gaps a candidate is least able to notice.
    """
    rows = []
    for c in components:
        looseness = (c["teachability"] - 1) / 4.0
        relevance = c["decision_relevance"] / 5.0
        rows.append({
            "id": c["id"],
            "looseness": round(looseness, 2),
            "decision_relevance": round(relevance, 2),
            "blind_spot_priority": round(looseness * relevance, 3),
```

```
        })
    return sorted(rows, key=lambda r: -r["blind_spot_priority"])

profile = context_profile(CONTEXT_COMPONENTS)
output_state["context_components"] = CONTEXT_COMPONENTS
output_state["context_profile"] = profile
output_state["top_blind_spot"] = profile[0]["id"]

# Normalise the decision-relevance ratings so that the two
    highest-weighted
# components carry the weights in later steps.
total = sum(c["decision_relevance"] for c in CONTEXT_COMPONENTS)
output_state["component_weights"] = {
    c["id"]: round(c["decision_relevance"] / total, 3) for c in
    CONTEXT_COMPONENTS
}
```

Reading the output. The ranking puts immersion-dependent and skill-dependent components at the top, which is the quantitative form of the intuitive claim that an outsider is least able to detect their own norms and strategy gaps. The honest use of this ranking is defensive: it tells you which of your own claims to hedge hardest. It does not tell you what to say.

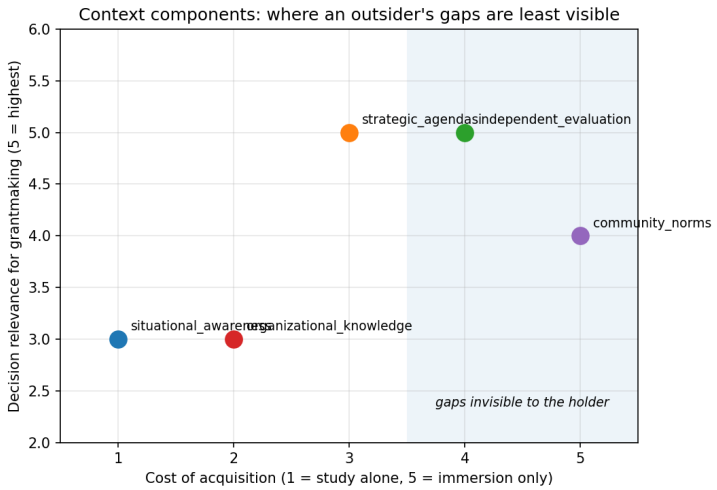


Figure 32: Graphical analysis for Step 1.

### Step 2

If component three is a skill, then in a live conversation context can only be **displayed, not asserted**. The corpus contains three registers of the same

requirement, and they are the same requirement, which is why an answer that satisfies one usually satisfies the others.

- Written register: D01 requires a linked claim-bearing summary, an indication of which considerations are most important, and a statement of confidence plus kind of support for major claims.
- Conversational register: P04 describes the habit of asking early and often what piece of evidence would be most likely to change one's mind, then seeking that evidence out with the cheapest possible check.
- Evaluative register: D04 lists reasoning transparency and scope sensitivity among the criteria by which applications are assessed.

The trap is that recitation of facts signals component one strongly while signalling the absence of component three. Reciting what has been read is the single most available conversational strategy and it targets the least weighted component. The mechanically safer move is to answer with a strategy, its premises, and its failure conditions.

The mechanical form of that move is a **what-must-be-true enumeration**. Take a strategy, list the premises it requires, and mark each premise as testable in principle or merely assumed, and as grounded in the corpus or not. The value is not the ratio; the value is that the untestable premises become visible before they are relied on, which is exactly the failure mode D01 was written to prevent.

```
# What-must-be-true enumerator.
# Converts a strategy claim into its premises, marks each premise as testable
# in principle or assumption-like, and flags whether the corpus supports it.
# This is the mechanical form of P09's third component and of D01's
# confidence-and-support requirement.

STRATEGY = {
    "name": "Fund a country-level AI safety fieldbuilding organisation",
    "premises": [
        {"claim": "a capable local pool exists but is not oriented to AI risk",
         "testable": True, "grounded_in_corpus": False,
         "check": "count of local AI and ML research staff"},
        {"claim": "catastrophic-risk framing motivates that pool",
         "testable": True, "grounded_in_corpus": False,
         "check": "prior programme conversion rates or a survey"},
        {"claim": "a small grant changes career decisions, i.e. money is the bottleneck",
         "testable": True, "grounded_in_corpus": True,
         "check": "P02 RFP volume; D03 and N01 rapid-grant behaviour"},
```

```
        {"claim": "the downstream transition rate into safety or governance work is high",
         "testable": True, "grounded_in_corpus": True,
         "check": "P07 HIP transition counts; P02 modelled SROI"},
        {"claim": "those transitions would not have happened anyway",
         "testable": False, "grounded_in_corpus": False,
         "check": "counterfactuality estimate, rarely measured"},
        {"claim": "the country's position creates distinctive leverage",
         "testable": False, "grounded_in_corpus": False,
         "check": "argument only; N03 concedes its own theory of change is not distinctive"},
    ],
}

def what_must_be_true(strategy):
    premises = strategy["premises"]
    n = len(premises)
    untestable = [p for p in premises if not p["testable"]]
    ungrounded = [p for p in premises if not p["grounded_in_corpus"]]
    untestable_share = len(untestable) / n
    return {
        "strategy": strategy["name"],
        "premise_count": n,
        "untestable_share": round(untestable_share, 2),
        "ungrounded_share": round(len(ungrounded) / n, 2),
        "load_bearing_and_untestable": [p["claim"] for p in untestable],
        "cheapest_useful_check": min(
            (p["check"] for p in premises if p["testable"] and not p["grounded_in_corpus"]),
            default=None,
        ),
        "verdict": (
            "state the untestable premises as assumptions and bound their effect on the decision"
            if untestable_share >= 0.3
            else "premises are mostly checkable; proceed to costing"
        ),
    }

output_state["what_must_be_true"] = what_must_be_true(STRATEGY)

# The same enumeration is reusable on someone else's proposal, which is the
# form it will take in an interview: the enumerator is run on the claim the
# interviewer offers, not only on one's own.
output_state["what_must_be_true_function_available"] = True
```

What this buys, and what it does not. Listing premises is not the same as evaluating them; the enumerator only makes the structure of a claim inspectable. Note also the connection to the earlier phase: the untestable premise in the example, counterfactuality, is the same multiplicative factor that the cost-effectiveness material insisted must be kept separate from the value estimate. The enumerator is where that separation becomes visible in prose.

**Step 3**

The limits of this account matter, and they should be stated rather than smoothed over. P09 is a programme report on a single four-day experiment with twenty-four participants, six facilitators and no control condition. It is not a validated instrument, we have no inter-rater reliability for **context**, and the source itself reports that participants needed different things and that the balance was not always right.

The corpus cannot tell us which components an interviewer weights. P09 defers that question explicitly. Any assertion of the form *Coefficient Giving evaluates candidates mainly on component two* is unsupported by these sources, even though the behavioural criteria in D04 are consistent with it.

Disposition. Use the five components as a self-orientation checklist and as shared vocabulary in an answer. Do not use them as a scoring rubric, and do not describe them as established measurement. The distinction between using a framework and mistaking it for a measurement is itself a norm in this community, and D01 is the document that says so.

## Sub-Question 2: Community norms: what they do, and which ones survive growth

Inventory the community's argumentative norms with documented anchors, explain what function each performs, and predict which norms are strained by the field's scale-up. Then isolate the pairwise tensions inside the norm set and name the earliest point of divergence in each.

**Step 1**

P12 supplies the load-bearing premise: the AI safety funding ecosystem started with a great deal of mutual trust because it was small, weird, demanding and socially interwoven, and this probably will not scale. P12 also cites the claim that trust is a bottleneck to growing teams quickly, and describes a principal-agent problem in which the party holding money cannot verify that the party holding taste is choosing well.

The non-trivial inference is this: if trust substitutes for verification, then a norm is not decoration but a **verification substitute**. A norm is a public prediction about behaviour in a case where enforcement is too costly: publish the reasoning even when the reasoning is unflattering; ask what would change

your mind before you defend your view; describe an opposing position in a form its holder would recognize. Each of those makes cooperation rational without requiring anyone to inspect anyone's motives. That is why the norms cluster around legibility rather than around loyalty.

The prediction this licenses is specific. Norms that can be converted into cheap public artifacts survive scaling, because conversion replaces unobservable shared context with something an outsider can check. Norms that depend on immersion degrade, because the new participants cannot acquire them by reading and the old participants cannot transmit them faster than they are hiring. P12's transparency programme, P10's commitment to quarterly public reports, and D01's demand that a reader be able to reconstruct the reasoning are all instances of the same conversion. P10's exception for work that should not be publicized is the refinement: the newest version of the norm distinguishes what must be published (the funder's reasoning about the grant) from what may not be published (the grantee's activity).

Table 61: Anchors are the documents that state or demonstrate the norm. The last column is inference about strain, not a source claim.

| Norm | What it requires | Anchor | Function it performs | What strains it |
|---|---|---|---|---|
| **Reasoning transparency** | Open with a linked summary; rank considerations; state confidence and kind of support | D01, D04 | Lets a reader update without having been in the room | Costly; L01 treats a vibes-based 1-to-7 KPI as acceptable, a weaker standard |
| **Public legibility of money** | Publish what was funded and why | P10, P12 | Builds trust at a distance and invites external audit | Private but impactful work (P10); credit allocation and double counting (S10) |
| **Crux-seeking** | Ask early what would change your mind, then look for it | P04, D01, D04 | Converts a disagreement into a testable question with a cost | Can become performative hedging if no decision follows |

| Norm | What it requires | Anchor | Function it performs | What strains it |
|---|---|---|---|---|
| **Funding people who disagree with you** | Support both sides of a live question | P08, D04 | Lets the best ideas win over time instead of the incumbent view | Competes with the urge to steer; P06 warns against the strategic-funder trap |
| **Nonpartisanship** | Keep catastrophic-risk work defensible across parties | P08, D04 | Preserves a coalition under two-party policymaking | S11 records a different complaint, that politics corrodes effective altruism itself |
| **Fast, low-friction response** | Reply in days, keep application cost low | D03, P13, P10, D04 | Respects the applicant's option value; some projects only exist if funded quickly | Volume; P04 warns against unnecessary work for organisations, D04 had to raise the bar on capacity grounds |
| **Scope sensitivity and explicit arithmetic** | Show rough magnitudes and what would change the decision | D04, D02 | Prevents category errors across grant sizes | L01 explicitly permits a vibes-based KPI where a number is unavailable |
| **Agency and ownership** | Take initiative on problems that nobody owns | Role brief, S02, S09 | Fills gaps that no job description covers | P06 warns ownership can slide into treating grantees as subcontractors |

```
import re

```

```
# Norm inventory with two ordinal axes.
# verifiability_cost: 1 = an outsider can check compliance in minutes,
#                     5 = only an insider can tell whether it was honoured.
# scaling_strain:     1 = survives growth roughly unchanged,
#                     5 = degrades quickly as the community grows.
NORMS = [
    {"id": "reasoning_transparency", "norm": "Reasoning transparency",
     "anchor": "D01; D04", "verifiability_cost": 4, "scaling_strain": 2},
    {"id": "public_money", "norm": "Public legibility of money",
     "anchor": "P10; P12", "verifiability_cost": 2, "scaling_strain": 2},
    {"id": "crux_seeking", "norm": "Crux-seeking before defending",
     "anchor": "P04; D01; D04", "verifiability_cost": 3, "scaling_strain": 3},
    {"id": "fund_dissent", "norm": "Funding people who disagree with you",
     "anchor": "P08; D04", "verifiability_cost": 3, "scaling_strain": 4},
    {"id": "nonpartisan", "norm": "Nonpartisanship",
     "anchor": "P08; D04", "verifiability_cost": 4, "scaling_strain": 4},
    {"id": "fast_response", "norm": "Fast, low-friction response",
     "anchor": "D03; P13; P10; D04", "verifiability_cost": 2, "scaling_strain": 5},
    {"id": "scope_sensitivity", "norm": "Scope sensitivity and explicit arithmetic",
     "anchor": "D04; D02", "verifiability_cost": 3, "scaling_strain": 2},
    {"id": "agency", "norm": "Agency and ownership",
     "anchor": "role brief; S02; S09", "verifiability_cost": 5, "scaling_strain": 3},
]

# Phrases that merge many actors into one voice. The brief may not use them
# unless a specific document is named as the holder of the view.
MERGED_VOICE_PATTERNS = [
    r"\bthe (ai safety|ea) (community|field|movement) (believes|thinks|agrees|wants|holds)\b",
    r"\beveryone (agrees|knows|believes|accepts)\b",
    r"\bthe field has (converged|settled)\b",
    r"\bcoefficient giving (thinks|believes|wants)\b",
    r"\bwe all know\b",
]
UNHEDGED_PATTERNS = [r"\b(certainly|obviously|clearly|undoubtedly|definitely)\b"]

def merged_voice_audit(text):
```

```
    flags = []
    lowered = text.lower()
    for pattern in MERGED_VOICE_PATTERNS + UNHEDGED_PATTERNS:
        for m in re.finditer(pattern, lowered):
            flags.append({"pattern": pattern, "match": m.group(0),
    "at": m.start()})
    return flags

SAMPLE = ("The AI safety community agrees that open weights are
    dangerous. "
          "Everyone knows that timelines are short, and it is obvious
    that "
          "Coefficient Giving thinks regulation is the answer.")

output_state["norm_inventory"] = NORMS
output_state["merged_voice_flags"] = merged_voice_audit(SAMPLE)

# Which norms are converted into checkable artifacts is the
    operational question
# after P12: conversion is how a tacit norm survives the growth P01
    and P06
# describe. Rank by (verifiability_cost + scaling_strain); low is
    durable.
output_state["durability_ranking"] = sorted(
    [{"id": n["id"], "durability_cost": n["verifiability_cost"] +
    n["scaling_strain"]}
     for n in NORMS],
    key=lambda r: r["durability_cost"],
)
```

Reading the ranking. Low combined cost means a norm is either cheap to check or robust to growth, or both. Public legibility of money and reasoning transparency come out durable. Agency and ownership comes out fragile, and P06's response to that fragility is instructive: rather than writing rules, the organisation says it will keep investing in ever more high-performance people, explicitly citing the practice of minimising rules and bureaucracy as an organisation scales. That is a bet that a fragile norm can be preserved by selection rather than by process, and it is a testable bet with a stated failure mode, since P06 also worries aloud about the pull toward risk aversion.

### Step 2

A norm inventory is not enough, because the norms conflict. Three pairwise tensions are visible in the corpus, and separating them from a fourth tension that is internal to one organisation is most of the work.

- Tension one. S11 wants maximal transparency and treats secrecy as bad in itself. P10 states that some impactful work requires privacy and that such grants will be reported in aggregate but not listed. The earliest point of

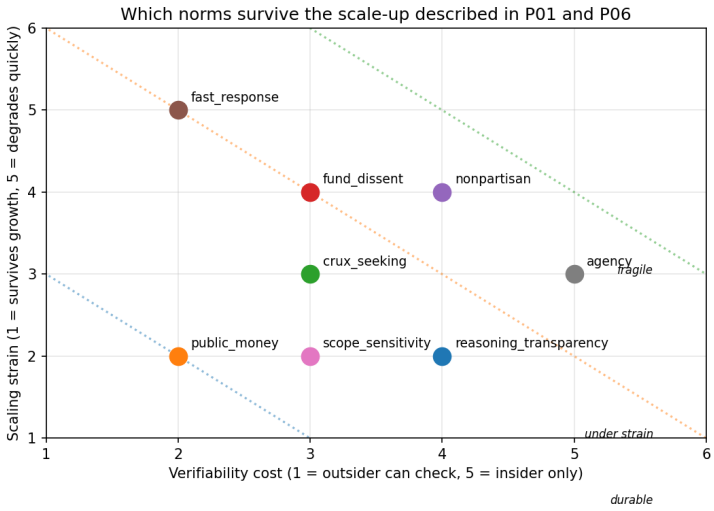


Figure 33: Graphical analysis for Step 1.

divergence is the **object of transparency**: the funder's reasoning, which can always be published, versus the grantee's activity, which sometimes cannot. Once the object is separated, most of the disagreement dissolves.

- Tension two. D01 recommends publishing data and code where possible. P08 lists information security for preventing unintended proliferation of dangerous systems among its safeguards, and D04 lists cybersecurity for model weights. The earliest divergence is whether the artifact's marginal value accrues to defenders or to attackers at that point on the capability curve. D01's data-and-code norm was formulated for research analysis, not for capability artifacts; the scope confusion is the disagreement.
- Tension three. P01 and S08 treat urgency as truthful and as the licence for speed, scale and occasional weirdness. S11 records that general fearmongering seems bad, and P08 argues that pulling safety into partisan frames is a mistake. The earliest divergence is empirical: whether the binding constraint on action is insufficient alarm or insufficient political feasibility. N14 is unusually direct evidence for the second, reporting that policy proposals were rejected for financial and political cost, and concluding that initiatives built on a majority expert view are not useful unless such a majority already exists.
- Tension four is internal to one organisation and should be labelled as such: P06 holds both that funders should own problems end to end like general managers, and that they should remain marginalists with the curiosity and humility of a funder open to grantee local knowledge. P06 says explicitly that the right balance will vary and that the organisation should

be intentional about striking it. Do not present this as a community-wide disagreement.

Two of the three cross-community tensions have the same shape: a norm is applied to an object outside its original scope. That is a recurring pattern and a cheap diagnostic. When a disagreement looks heated but unresolvable, ask which object the norm was written for, and whether the object under discussion is the same one.

```
def earliest_divergence(chain_a, chain_b):
    """Longest common prefix of two stated premise chains.

    Each chain is ordered from shared background towards the claim.
    The last
    element of the common prefix is the deepest premise the two
    positions
    still share; the next elements are the first substantive
    disagreement.

    Caveat: real arguments are directed acyclic graphs, not chains.
    Treat this
    as a disciplined prompt for locating the first divergence, not as
    a proof
    that the divergence is unique. Re-ordering or inserting an
    implicit
    premise moves the reported point, so state the chains explicitly.
    """
    i = 0
    while i < min(len(chain_a), len(chain_b)) and chain_a[i] ==
    chain_b[i]:
        i += 1
    shared = chain_a[:i]
    return {
        "shared_prefix": shared,
        "last_shared_premise": shared[-1] if shared else None,
        "divergence_index": i,
        "position_a_next": chain_a[i] if i < len(chain_a) else None,
        "position_b_next": chain_b[i] if i < len(chain_b) else None,
    }

TENSIONS = {
    "transparency_vs_privacy": {
        "a": ["funders should be legible",
              "legibility means publishing what was funded",
              "therefore publish grantee identities"],
        "b": ["funders should be legible",
              "legibility means publishing the reasoning",
              "some grantee activity is harmed by disclosure",
              "therefore publish reasoning and aggregate the rest"],
    },
```

```
    "openness_vs_security": {
        "a": ["research artifacts should be reproducible",
              "reproducibility requires data and code",
              "therefore publish artifacts"],
        "b": ["research artifacts should be reproducible",
              "some artifacts are capability transfers",
              "reproducibility is not the relevant property for those",
              "therefore publish selectively"],
    },
    "urgency_vs_feasibility": {
        "a": ["the risk is severe",
              "severe risk justifies alarm",
              "alarm accelerates action",
              "therefore raise the alarm"],
        "b": ["the risk is severe",
              "policy requires a durable coalition",
              "alarm narrows the coalition",
              "therefore build feasibility first"],
    },
}

output_state["tension_divergences"] = {
    name: earliest_divergence(t["a"], t["b"]) for name, t in TENSIONS.items()
}

# The pattern to look for: whether the split is a scope error about the object
# a norm applies to, an empirical premise, or a values premise. Only the
# middle kind is worth a costly check.
output_state["divergence_kinds"] = {
    "transparency_vs_privacy": "scope error about the object of transparency",
    "openness_vs_security": "scope error about the object of the artifact",
    "urgency_vs_feasibility": "empirical premise about the binding constraint",
}
```

What this yields. Two of the three divergences are scope errors, and the correct response to a scope error is a definitional clarification rather than an investigation. Only the third is empirical and worth a check. This is a useful asymmetry: most apparent community conflicts are cheaper to dissolve by making the object explicit than to settle by evidence, and the ability to tell which kind you are facing is a compact demonstration of context.

## Sub-Question 3: The disagreement map, with cruxes preserved

Build the map: eight live disagreements, each stated with two defensible positions, the earliest point of divergence, and the evidence that would move it. Then rank the cruxes by expected value of information and select what to check inside a fixed time budget.

### Step 1

An entry in the map must satisfy five conditions. It must state each side in a form its holder would recognize; it must name the earliest point of divergence rather than the loudest; it must name evidence that would move at least one side; it must record whether that evidence changes the recommendation or only the confidence; and it must record what is not in dispute.

The underlying discipline comes from L02's third exercise: prepare both sides of a consequential discussion, locate the earliest point of disagreement, specify what evidence or clarification would bear on that point, and revise your own argument when the comparison requires it. The crux definition comes from the previous phase: a possible fact that is uncertain, decision-relevant, and resolvable at reasonable cost. A risk is not a crux, and an unsettled value is not a crux either.

Two disciplines are required throughout. First, the anti-merging rule: attribute every position to a document, an author, or a group a document identifies. Second, the internal-external distinction: P06's marginalism-versus-ownership tension is inside one funder, P08's disagreement-funding policy is an institutional position, and S11's list is one person's informal sketch. Flattening these three into a field-wide debate is the single most likely error in this exercise.

### Step 2

Table 62: Positions are attributed to sources, not to the field. Rows 1, 2, 3, 5, 7 and 8 are cross-community; row 6 is internal to one funder; row 4 carries both an institutional position and one informal dissent.

| Disagreement | One defensible position | Opposing defensible position | Earliest point of divergence | Evidence that would move it |
|---|---|---|---|---|
| **1. Openness of model weights** | Open release is on net good: capability diffuses anyway, and access raises defensive and verification capacity (S11) | Release is an irreversible transfer; misuse uplift is real and concentrated control is the better lever (P08, D04) | The size of marginal misuse uplift from release relative to marginal defensive uplift, at a given position on the capability curve | Hazardous-capability evaluations (WMDP-style, P11); uplift studies; post-release incident tracking; whether safety training survives fine-tuning |
| **2. Pace, pause, or default** | We may already be in a pause or pace-by-default world, so the dramatic frame is stale (S11) | Progress has repeatedly outpaced expectations and short-timeline predictors have been more accurate, so plan for speed (P01, P06) | Whether any binding decelerating mechanism is actually operating, as opposed to the pace simply continuing | Dated capability and time-horizon series; compute and data-centre capital commitments (P08); regulatory instruments actually in force |

| Disagreement | One defensible position | Opposing defensible position | Earliest point of divergence | Evidence that would move it |
|---|---|---|---|---|
| **3. Timelines** | Short: transformative AI plausibly within a few years (P01) | Medium or broad (S11) | The relative weight of capability scaling versus diffusion and deployment drag; P03 spans five to thirty years with a central guess of fifteen | Agent task-length doubling every four to seven months (P03); whether domain experts keep beating superforecasters on capability questions (P06) |
| **4. Geopolitical framing** | Democracies should maintain a competitive advantage in AI over authoritarian states (P08) | The fear of China is not understood, and politics is bad for the soul of effective altruism (S11) | Whether a capability lead is durable enough to convert into a governance advantage, and whether racing or diffusion is the binding risk | Effectiveness of export controls; supply-chain chokepoints; the measured size and decay rate of capability and adoption leads |

| Disagreement | One defensible position | Opposing defensible position | Earliest point of divergence | Evidence that would move it |
|---|---|---|---|---|
| **5. The binding constraint as funding scales** | Organisations are scarce and money will be absorbed by organisations that can hire and run projects (P10) | Grantmaker staff time is the constraint, as the funder itself reports (P01); and trust is the constraint on the coordination layer (P12) | The choice of denominator: grants per grantmaker-hour, organisations per dollar, or verified agent pairs per unit of coordination cost | Application volume and conversion (P02 expected under a 25 percent chance of more than forty applications and received 474); investigations per grantmaker, six to twelve (P04); D04 raising the bar on capacity grounds |
| **6. What a funder should be** | Own problems end to end like a general manager; seed organisations and recruit founders (P06, S09) | Remain marginalist, keeping the curiosity and humility of a funder open to grantee local knowledge (P06) | Internal to one funder and explicitly unresolved there. The empirical question is which produces more expected value per grantmaker-hour, and at what horizon | Track record of funder-initiated organisations; cost per dollar moved for existing versus new grantees; P06's note that seeding is more labour intensive per dollar early |

| Disagreement | One defensible position | Opposing defensible position | Earliest point of divergence | Evidence that would move it |
|---|---|---|---|---|
| **7. Urgency framing** | Urgency is truthful and licenses speed, scale and occasional weirdness (P01, S08) | Fearmongering is bad; partisan framing narrows the coalition (S11, P08) | Whether the binding constraint on action is insufficient alarm or insufficient political feasibility | N14's account of UK policy engagement: proposals were rejected for financial and political cost, and initiatives resting on an expert majority are useless without one |
| **8. Funding mechanism and credit** | A grant-only donation model keeps the institution credible and the grantee incentives clean | Fund on value delivered rather than cost, consider equity-like instruments to recycle upside, and shorten response-time games between funders (S10, P16, P17) | Whether equity-like instruments increase total capital available without distorting grantee selection, and whether credit and double counting can be specified well enough to compare funders | P16's SAFE experiment and its opt-out clause; P17's unresolved treatment of a nonprofit that later incorporates; P10's power-law payoff model |

What is not in dispute, and therefore must not be presented as contested. That capabilities are advancing; that catastrophic risk from advanced AI is plausible enough to justify funding; that governance mechanisms matter at the margin; that measurement and evaluation quality are valuable; and that the field is underfunded relative to the opportunity set, which P08 states directly. Row five's trust position does not contradict the other two positions in that row: it sits at a different layer of the same system, and the earliest divergence there is the

choice of denominator rather than a factual dispute about magnitudes. Rows one through four are genuine factual disagreements with shared background premises.

```
import math

# Each crux is scored on three axes so the map can be prioritised without
# pretending the axes are commensurable.
#   uncertainty: how split informed actors are, 0 = settled, 1 = maximally split
#   decisiveness: does resolving it change the recommendation or only the
#                 confidence, 0 = neither, 1 = flips the decision
#   cost_hours: order-of-magnitude cost of the cheapest useful check
CRUXES = [
    {"id": "open_weights_uplift", "uncertainty": 0.7, "decisiveness": 0.8, "cost_hours": 40},
    {"id": "pace_mechanism", "uncertainty": 0.6, "decisiveness": 0.5, "cost_hours": 8},
    {"id": "timeline_weighting", "uncertainty": 0.8, "decisiveness": 0.7, "cost_hours": 20},
    {"id": "capability_lead_durability", "uncertainty": 0.7, "decisiveness": 0.4, "cost_hours": 30},
    {"id": "binding_constraint_denominator", "uncertainty": 0.5, "decisiveness": 0.9, "cost_hours": 6},
    {"id": "ownership_versus_marginalism", "uncertainty": 0.8, "decisiveness": 0.6, "cost_hours": 80},
    {"id": "urgency_versus_feasibility", "uncertainty": 0.5, "decisiveness": 0.6, "cost_hours": 12},
    {"id": "equity_instrument_effect", "uncertainty": 0.6, "decisiveness": 0.5, "cost_hours": 25},
]

def crux_priority(c):
    """Priority = uncertainty x decisiveness / log2(1 + cost in hours).

    The log keeps a costly check from being penalised into irrelevance; it
    does not make the score a ratio of anything real. Use the ordering only,
    and say so when reporting it.
    """
    return c["uncertainty"] * c["decisiveness"] / math.log2(1 + c["cost_hours"])

ranked = sorted(CRUXES, key=crux_priority, reverse=True)
for row in ranked:
    row["priority"] = round(crux_priority(row), 3)
```

```
output_state["crux_ranking"] = ranked
output_state["top_crux"] = ranked[0]["id"]
```

Reading the ranking, and its limits. The inexpensive structural question about the denominator outranks the expensive factual question about weight release, which is exactly the asymmetry a grantmaker on a two-day investigation should exploit: a denominator disagreement can be dissolved by definitional work, while an uplift question needs a study that does not exist yet. Two caveats. First, the axis values are judgments and the ranking is ordinal only. Second, decisiveness is not stable across contexts: the same crux can be decisive for a small grant and irrelevant for a large one, because the cost of being wrong scales with the ticket.

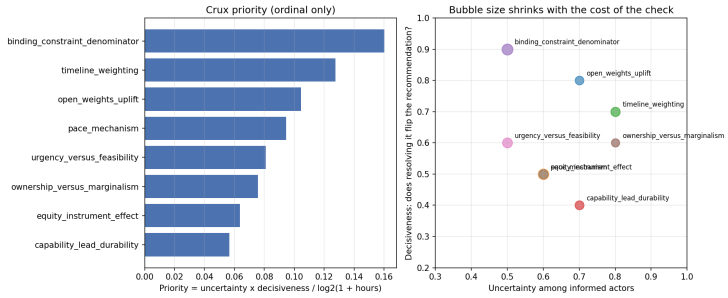


Figure 34: Graphical analysis for Step 2.

### Step 3

Priority scores do not answer the operative question, which is what to do with the hours actually available. A grantmaker carries several investigations at once (P04 reports six to twelve), so the relevant problem is selection under a budget rather than ranking in the abstract.

Two tie-break rules make the selection defensible. First, prefer a check whose result you can act on within the current decision window; a perfect study that lands after the money is committed has no decision value. Second, prefer a check that also downgrades a premise you were relying on silently; a check that can only confirm is a poor use of an hour.

```
import math

# Greedy selection of crux checks under a fixed hour budget.
# Each check has a priority and a cost; some are prerequisites for others,
# which is the only structure that makes a greedy pass insufficient. Where a
```

```
# prerequisite exists, the prerequisite is charged to the same budget.
CHECKS = [
    {"id": "define_denominator", "priority": 0.075, "hours": 6, "needs": []},
    {"id": "timeline_series_read", "priority": 0.078, "hours": 20, "needs": []},
    {"id": "pace_mechanism_scan", "priority": 0.083, "hours": 8, "needs": []},
    {"id": "urgency_feasibility_interview", "priority": 0.058, "hours": 12, "needs": []},
    {"id": "uplift_literature_read", "priority": 0.049, "hours": 40, "needs": ["define_denominator"]},
    {"id": "ownership_track_record", "priority": 0.040, "hours": 80, "needs": []},
]

def select_checks(checks, budget_hours):
    by_id = {c["id"]: c for c in checks}
    chosen, spent, skipped = [], 0, []
    for c in sorted(checks, key=lambda x: -x["priority"]):
        chain = []
        cursor = c
        seen = set()
        while cursor and cursor["id"] not in seen:
            seen.add(cursor["id"])
            chain.append(cursor)
            cursor = by_id.get(cursor["needs"][0]) if cursor["needs"] else None
        chain_cost = sum(step["hours"] for step in chain if step["id"] not in [x["id"] for x in chosen])
        if spent + chain_cost <= budget_hours:
            for step in reversed(chain):
                if step["id"] not in [x["id"] for x in chosen]:
                    chosen.append(step)
            spent += chain_cost
        else:
            skipped.append({"id": c["id"], "reason": "budget", "chain_cost": chain_cost})
    return {"chosen": chosen, "hours_spent": spent, "skipped": skipped}

output_state["selection_48h"] = select_checks(CHECKS, 48)
output_state["selection_90min"] = select_checks(CHECKS, 1.5)
```

Why this belongs in a norms brief. The output is a schedule, and the schedule is itself an argument: it states which disagreements are cheap to dissolve, which are cheap to bound, and which are expensive to resolve and should therefore be treated as assumptions with bounded effect. In a conversation, that structure is what distinguishes someone who knows the field from someone who has read

about it. The number is scaffolding; the ordering is the claim.

## Sub-Question 4: The earliest point of disagreement: a worked exercise with two defensible answers

Apply the earliest-divergence method to one live disagreement, construct two defensible answers each with its assumption, decisive consideration and reversal condition, and write the calibrated interview script that presents them without merging voices.

### Step 1

Worked case: the openness of model weights. The trap is that this question looks like a values dispute and is usually treated as an empirical one, or the reverse. Decomposing the two positions into premise chains makes the confusion visible.

Position A, drawn from S11: capabilities advance regardless of release policy; defensive and verification capacity scale with access to models; concentrated control over capable systems is itself a catastrophic risk; therefore the marginal misuse uplift from release is small relative to the marginal defensive uplift, and release is preferred. Position B, drawn from P08 and D04: capabilities advance regardless of release policy; misuse uplift from widely available weights is real and non-trivial, including for biological and cyberoffence capabilities; concentration of control carries risk but the correct lever is governance rather than openness; therefore release should be restricted, at least at the frontier.

The shared prefix is one premise, that capability advance is largely independent of any single actor's release policy. The first substantive divergence is the second premise. This is a large result for a small amount of work: the disagreement is mostly empirical, it is located at the claim about marginal uplift, and it is therefore addressable by measurement rather than by negotiation. P08 itself concedes the crux is open, quoting an earlier staff post that asks whether tighter regulation would reduce catastrophic risk or increase it by privileging development in jurisdictions with lower safety standards and less cooperation. That is the same crux stated from the other direction, which is unusually strong evidence that the funder holds it open rather than settled.

Second application, to row five of the map. The three constraint claims do not diverge over a fact about the world; they diverge over the choice of denominator. Grants per grantmaker-hour, organisations per dollar, and verified agent pairs per unit of coordination cost are three different production functions, and a disagreement between them cannot be settled by a measurement because the parties are measuring different things. The correct first move is definitional.

```
def earliest_divergence(chain_a, chain_b):
    i = 0
    while i < min(len(chain_a), len(chain_b)) and chain_a[i] ==
    chain_b[i]:
```

```
        i += 1
    shared = chain_a[:i]
    return {
        "shared_prefix": shared,
        "last_shared_premise": shared[-1] if shared else None,
        "divergence_index": i,
        "position_a_next": chain_a[i] if i < len(chain_a) else None,
        "position_b_next": chain_b[i] if i < len(chain_b) else None,
    }

OPEN_WEIGHTS_A = [
    "capabilities advance largely independently of any single release policy",
    "defensive and verification capacity scale with access to models",
    "concentrated control is itself a catastrophic risk",
    "marginal misuse uplift is small relative to marginal defensive uplift",
    "therefore default towards release",
]
OPEN_WEIGHTS_B = [
    "capabilities advance largely independently of any single release policy",
    "misuse uplift from widely available weights is real and non-trivial",
    "concentration of control carries risk but governance is the lever",
    "marginal misuse uplift exceeds marginal defensive uplift at the frontier",
    "therefore restrict release at the frontier",
]

CONSTRAINT_ORGS = [
    "a large funding wave is expected",
    "money is absorbed by organisations that can hire and run projects",
    "organisations per dollar is the binding ratio",
]
CONSTRAINT_STAFF = [
    "a large funding wave is expected",
    "grantmaking throughput is limited by grantmaker attention",
    "grants per grantmaker-hour is the binding ratio",
]
CONSTRAINT_TRUST = [
    "a large funding wave is expected",
    "deployment requires matching money to unverified judgment",
    "verified agent pairs per unit of coordination cost is the binding ratio",
]

output_state["open_weights_divergence"] =
```

```
    earliest_divergence(OPEN_WEIGHTS_A, OPEN_WEIGHTS_B)
output_state["constraint_divergences"] = {
    "orgs_vs_staff": earliest_divergence(CONSTRAINT_ORGS,
    CONSTRAINT_STAFF),
    "staff_vs_trust": earliest_divergence(CONSTRAINT_STAFF,
    CONSTRAINT_TRUST),
    "orgs_vs_trust": earliest_divergence(CONSTRAINT_ORGS,
    CONSTRAINT_TRUST),
}
output_state["divergence_type_open_weights"] = "empirical premise:
    size of marginal misuse uplift relative to marginal defensive
    uplift"
output_state["divergence_type_constraint"] = "definitional: choice of
    denominator, not a fact about magnitudes"
```

Methodological caveat. The chains are stated by us, and chain order is doing work. Inserting an implicit premise, such as the assumption that a released weight cannot be recalled, moves the reported divergence point one step earlier. Say this out loud when using the method: the reported divergence is a function of the stated chains, and the discipline is to state them so they can be argued with.

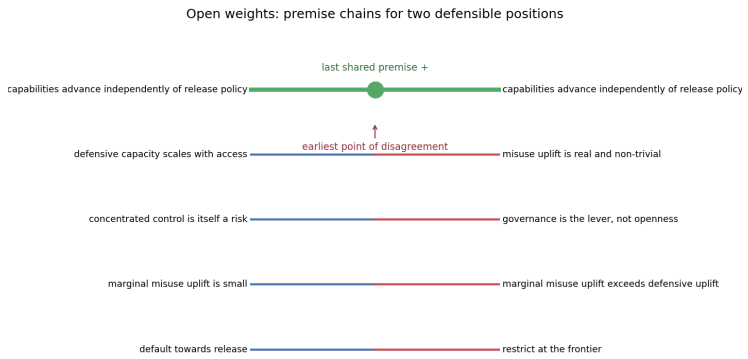


Figure 35: Graphical analysis for Step 1.

### Step 2

The two answers below are both defensible on the supplied corpus. Neither is the community's answer, because there is no such thing. What makes them useful is that each names the assumption it rests on, the consideration that decides it, and the condition that reverses it.

Table 63: Both answers share the premise that capabilities advance independently of release policy, which is why the disagreement is about uplift magnitude and irreversibility rather than about whether progress happens.

| Element | Answer one: graduated release | Answer two: restricted release pending measurement |
|---|---|---|
| Headline recommendation | Release below the frontier, restrict at and above it, and tie the threshold to evaluated hazardous capability rather than parameter count | Default to restricted release at the frontier, with pre-committed review points that convert the default once measurements exist |
| Core assumption | Misuse uplift rises with capability faster than defensive uplift, so a capability threshold exists that is worth defending | Irreversibility dominates: a released weight cannot be recalled, so the option value of waiting exceeds the expected value of diffuse defensive use |
| Decisive consideration | The differential uplift measurement at the margin, which is exactly the crux the two positions split on | Asymmetry of consequences under uncertainty, not the point estimate of uplift |
| What would change it | Evidence that uplift is flat across the capability curve, which collapses the graduated threshold into uniform openness; or evidence that defensive uses depend on frontier-adjacent weights, which pushes the threshold outward | Evidence that restriction measurably degrades defensive capacity, or that it displaces development towards actors with weaker safety practices, which is the failure mode named in P08's own earlier statement of the crux |

| Element | Answer one: graduated release | Answer two: restricted release pending measurement |
|---|---|---|
| Residual risk if wrong | A capability threshold is mis-set and a hazardous model is released; the error is detectable and partially correctable, since release is reversible in policy even when weights are not | Restriction buys time that is then spent without producing the measurement, and the field loses defensive capacity it will need later; the error is slow to detect |

How this would be judged. The strength is in the structure, not in which column is chosen: each answer states its assumption, names a decisive consideration, and gives a reversal condition that could actually occur. D04's assessment criteria reward exactly this pattern, asking for strategic judgment under uncertainty, reasoning transparency and scope sensitivity, and asking applicants to identify failure modes and prioritise them rather than listing every risk. An answer that picks one column and defends it without a reversal condition is weaker than one that picks either column with the conditions attached. Note also that both answers are consistent with funding people who disagree, which is the institutional position in P08, so endorsing one position does not require treating the other as bad faith.

**Step 3**

The script below has six fixed slots. They are the conversational form of D01's three recommendations plus L02's methodology move, and P04's habit of naming the evidence that would change one's mind before looking for it.

1. State what is not in dispute, so the disagreement is located rather than implied.
2. Give a position with a probability and name the kind of support it rests on.
3. Name the earliest point of divergence and classify it: empirical, definitional, or values.
4. Name the cheapest check that would move the crux and its rough cost.
5. State the reversal condition explicitly.
6. Flag where the support is thin and what you would need to learn.

The calibration step is mechanical. Convert the prior probability to log-odds, add the log-likelihood ratio implied by the observation, and convert back. This

keeps the update auditable and makes the strength of the evidence visible instead of hiding it in a revised adjective.

$$\log \frac{p_1}{1-p_1} = \log \frac{p_0}{1-p_0} + \log \Lambda$$

$$p_1 = \left(1 + \exp\left(-\left(\log \frac{p_0}{1-p_0} + \log \Lambda\right)\right)\right)^{-1}$$

```
import math

# Six fixed slots for a calibrated answer on a contested question.
def calibration(p0, likelihood_ratio):
    log_odds_0 = math.log(p0 / (1 - p0))
    log_odds_1 = log_odds_0 + math.log(likelihood_ratio)
    p1 = 1 / (1 + math.exp(-log_odds_1))
    return {"prior": round(p0, 3), "likelihood_ratio":
    likelihood_ratio,
            "posterior": round(p1, 3), "log_odds_shift_bits":
    round((log_odds_1 - log_odds_0) / math.log(2), 2)}

SLOTS = {
    "01_not_in_dispute": (
        "Both positions agree that capability progress is largely
    independent of any "
        "single release decision, and that catastrophic misuse risk
    is real enough to "
        "warrant resources. The disagreement is not about whether to
    take the risk "
        "seriously."
    ),
    "02_position_with_confidence": (
        "I would put roughly 55 to 65 percent on graduated release
    being better than "
        "a frontier-wide restriction over the next few years. Support
    kind: a simple "
        "argument from option value plus the observation that
    defensive work is "
        "measurement-limited today. That is a weak-support claim, not
    a studied one."
    ),
    "03_earliest_divergence": (
        "The divergence is at the second premise, the size of
    marginal misuse uplift "
        "relative to marginal defensive uplift. It is an empirical
    premise rather than "
        "a values premise, so it can in principle be measured rather
    than negotiated."
    ),
```

```
    "04_cheapest_check": (
        "The cheapest useful check is hazardous-capability and uplift evaluation on "
        "models near the release threshold, of the kind the benchmark literature has "
        "started to build. It is not cheap in absolute terms, on the order of weeks of "
        "specialist time, but it is far cheaper than resolving the question by "
        "experience."
    ),
    "05_reversal_condition": (
        "I would move towards restriction if restriction were shown to degrade "
        "defensive capacity materially, or if capability development demonstrably "
        "shifted towards actors with weaker safety practices as a result. Both of "
        "those are the failure modes named in the funder's own earlier framing of "
        "the question, which is why I flag them rather than my own."
    ),
    "06_thin_support": (
        "Thin support on two things: my prior for uplift scaling with capability is "
        "an intuition, and I have no independent estimate of how quickly safety "
        "training degrades under fine-tuning. I would want to know both before "
        "raising my confidence above the mid-fifties."
    ),
}

output_state["answer_slots"] = SLOTS
output_state["prior_update_demo"] = {
    "weak_evidence": calibration(0.60, 1.5),
    "moderate_evidence": calibration(0.60, 5.0),
    "strong_evidence": calibration(0.60, 40.0),
}

# Anti-merging guard: any slot that asserts a collective belief must name a
# document. This is the operational form of the constraint on this brief.
output_state["merging_guard"] = (
    "Every claim in the slots is attributed to a position or to an argument, "
    "never to a community. No personal history, membership or credential is "
    "asserted anywhere in the script."
```

```
)
```

Why the calibration step earns its place. A likelihood ratio of forty is strong enough to move a coin-flip prior to the high nineties; a ratio of one and a half is noise. Writing both down forces the speaker to distinguish having read something from having learned something, which is the distinction D01 asks a writer to make about kinds of support and which L01 concedes is often abandoned, since it permits a KPI that is literally a one-to-seven guess. The interviewer's payoff is not the number; it is the visible refusal to treat a weak check as a strong one.

Closing observation for the norms brief as a whole. The three registers of the same requirement (D01's written discipline, P04's conversational habit, D04's assessment criteria) converge on one behaviour: make the crux visible, name what would move it, and state what you do and do not have. On contested terrain, where the community deliberately funds people who disagree, that behaviour is the only durable basis for a shared judgment. It is also the safest thing a candidate can demonstrate, because it depends on no claim about personal experience and survives any disagreement about which position is correct.

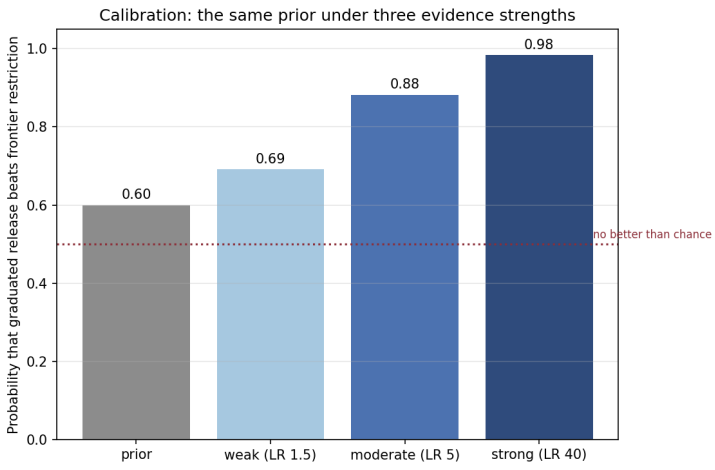


Figure 36: Graphical analysis for Step 3.

# Build a biosecurity and pandemic-preparedness brief.

## Question

Active step: build a biosecurity and pandemic-preparedness brief. The deliverable must cover biological risk pathways, CG's four-pillar plan, talent bottlenecks, example grants, and disputed risk estimates, and it must be usable by a generalist evaluator who may not be a wet-lab biologist.

*The brief should separate source-supported claims from inferences and proposed exercises, preserve uncertainty about biorisk estimates and intervention effectiveness, and connect each recommendation to a source.*

## Background

Source anchors: P01 reports CG biosecurity grantmaking growing from $30m in 2024 and $65m in 2025 to more than $150m in 2026. P04 describes Chris Bakerlee's evolutionary-biology path, field building, investigating 6-12 grants, RFPs, proactive individual grants, and the ERA AIxBio Research Fellowship. P06 names Andrew Snyder-Beattie, mirror bacteria, and a four-pillared plan: personal protective equipment such as elastomeric respirators; biohardening of buildings and infrastructure; pathogen-agnostic detection systems; and rapid medical countermeasures.

Additional anchors: P07 suggests Cures Within Reach for drug repurposing and discusses the market failure from lack of exclusivity. S04 describes the Strep A Vaccine Fund: over 600k deaths per year, about $14m per year of R&D globally before the fund, about 50 times less than malaria which has similar mortality, more than $140m raised at launch, and a goal to double clinical candidates and have at least one ready for Phase 3 by the end of 2030. N14 describes the pitfalls of an IPCC-like biotechnology risk body, including absent consensus, misaligned incentives and lack of willing champions. P08 records CG funding both a paper reporting LMs could help people with little or no lab training develop new pathogens and a RAND paper finding current LMs did not substantially increase biological attack risk.

*Uncertainty to preserve: biorisk estimates and intervention effectiveness are contested. The corpus does not supply base rates for natural pandemics, accidental releases, or deliberate attacks, and it does not supply validated effect sizes for biohardening, pathogen-agnostic detection, or rapid medical countermeasures. Any brief that states a single probability for biological catastrophe without a source is an invention.*

## Sub-Question 1: Biological Risk Pathways and the Distinctions an Evaluator Must Preserve

What are the major biological risk pathways, and which distinctions should an evaluator preserve before judging an intervention?

### Step 1

Biological risk is not one problem. The first evaluator move is to classify the pathway before judging the intervention. A grant that reduces the probability of a natural pandemic is not substitutable for a grant that reduces the probability of a deliberate attack, and neither is substitutable for a grant that improves response if a release occurs.

- Biosafety versus biosecurity versus biodefense versus global health security. Biosafety addresses accidents; biosecurity addresses deliberate misuse; biodefense addresses state and military response; global health security addresses natural outbreaks and health systems.
- Natural versus accidental versus deliberate. The mechanism, the relevant actors, the evidence base, and the intervention points differ across these three.
- State versus non-state actor. A state biological weapons program raises different governance, deterrence and verification questions than a terrorist or lone-actor pathway.
- Engineered versus naturally evolving pathogen. Engineering changes the plausibility of specific countermeasures and the relevance of synthesis screening and dual-use research oversight.
- AI-enabled capability uplift versus access barrier reduction. P08's two funded papers disagree about whether current language models substantially increase biological attack risk. One mechanism is uplift; another is lowering access and knowledge barriers.
- Mirror bacteria as a separate category. P06 names Andrew Snyder-Beattie's work on reducing risks from mirror bacteria. This is a synthetic-biology pathway whose risk profile is distinct from pathogen engineering or natural spillover.

Table 64: Table 1. Biological risk pathway matrix. The final column is the decision-relevant inference.

| **Pathway family** | **Mechanism** | **Source anchor** | **Principal uncertainty** | **Who disputes it** | **Grantmaker relevance** |
|---|---|---|---|---|---|
| Natural pandemic | Zoonotic spillover and natural evolution | P04 mentions pandemic preparedness; the corpus gives no quantified base rate | Frequency, severity, and speed of detection | Not specified in corpus | Global health security and pandemic preparedness are adjacent portfolios |
| Accidental release | Laboratory accident, biosafety failure, or containment breach | P04 describes wet-lab research; N14 discusses biotech risk governance | Base rate, containment quality, and reporting | Not specified in corpus | Biohardening, PPE, and detection grants act after release or at release |
| Deliberate misuse | State or non-state actor develops or uses a biological weapon | P08 records CG funding both sides of the AI-bio uplift debate | Actor intent, technical capability, and access to materials | The two funded P08 papers point in opposite directions | Governance, screening, and deterrence are prevention-side; most four-pillar grants are response-side |

| Pathway family | Mechanism | Source anchor | Principal uncertainty | Who disputes it | Grantmaker relevance |
|---|---|---|---|---|---|
| AI-enabled bio | Language models or agents lower the knowledge and design barrier | P08 reports one paper finding uplift for novices and one RAND paper finding no substantial increase in attack risk | Scaffolding, model access, and real-world execution bottlenecks | The two P08-funded research groups disagree | A live crux for AI-bio grantmaking; fund evidence rather than assume a side |
| Mirror bacteria | Synthetic organisms with mirrored chirality evade natural predators and immune systems | P06 names mirror bacteria and Andrew Snyder-Beattie's work on the risk | Feasibility, containment, and whether the risk is near-term | Not specified in corpus | A novel x-risk category; evaluate as a separate pathway, not as generic biorisk |

*Nontrivial insight: prevention, detection, response, and recovery are different intervention points. CG's four-pillar plan acts mostly at detection and response, not prevention. Prevention lives in governance, synthesis screening, biological weapons convention norms, and dual-use research oversight. An evaluator should ask which point in the chain the grant moves and what it assumes about the other points.*

$$R = P(\text{release}) \times P(\text{amplification}) \times I$$

*Caveat: the decomposition above is a proposed exercise, not a source-supported model. The corpus does not provide parameter values, and different pathways require different decompositions.*

## Sub-Question 2: CG's Biosecurity Strategy Compared with AI Safety and Global Health

What is CG's biosecurity strategy, and how does it compare with AI safety and global health?

**Step 1**

Table 65: Table 2. Funding scale across portfolios. Figures are source-reported and use different definitions; do not compare naively.

| Portfolio | 2024 | 2025 | 2026 so far or target | Source |
|---|---|---|---|---|
| Biosecurity and pandemic preparedness | $30m | $65m | more than $150m | P01 |
| Technical AI safety, security, and fieldbuilding | $168m | $351m | on track for over $1b | P01 |
| Global health and wellbeing | not broken out in P01 | part of over $1b total | S01 records a $212m bed-net grant | P06, S01 |

Strategic features of CG's biosecurity work: field building through fellowships and programs; a recent RFP; proactive individual transition grants; the four-pillar plan; work on mirror bacteria; and a general-manager style of ownership exemplified by Andrew Snyder-Beattie. P04 shows the work is varied: investigating 6-12 grants at a time, running an RFP, and making grants that would not have happened without active scouting.

- Personal protective equipment such as elastomeric respirators.
- Biohardening of buildings and infrastructure.
- Pathogen-agnostic detection systems.
- Rapid medical countermeasures.

*Nontrivial comparison: the biosecurity strategy is more resilience-and-response-focused, while the AI safety strategy is more prevention, control, and governance focused. Both share field-building and talent bottlenecks. But biosecurity has fewer dollars and fewer grants, so each grant carries higher variance and the evaluator needs either domain expertise or a disciplined deference protocol.*

$$\frac{150}{1000} \approx 0.15$$

Interpretation: even on the highest reported biosecurity figure, the portfolio is roughly an order of magnitude smaller than the AI safety target. That makes marginal biosecurity dollars potentially more neglected, but it also means the evaluation infrastructure and evidence base are thinner. The generalist should not treat smaller as automatically higher leverage.

*Caveat: P01 notes that committed grants approximate money reasonably sure to reach grantees and exclude conditionally approved grants. The ratios above are illustrative, not a CG institutional ranking.*

## Sub-Question 3: A Biosecurity Grant Evaluation Rubric and Three Worked Examples

How should an evaluator assess a biosecurity grant? Build a rubric and apply it to the ERA AIxBio Research Fellowship, Cures Within Reach, and the Strep A Vaccine Fund.

### Step 1

Table 66: Table 3. Biosecurity grant evaluation rubric. The rubric is an inference from P04, P07, S04, N14, and P08.

| **Criterion** | **Question** | **Evidence to seek** | **Red flags** | **Analogous AI safety question** |
|---|---|---|---|---|
| Theory of change | Does the grant plausibly reduce a catastrophic biological risk? | Causal chain from output to risk reduction | Outputs that do not connect to risk | Does the project reduce worst-case AI risk? |
| Neglectedness | Is this opportunity underfunded relative to importance? | Existing funders, R&D spend, field size | Crowding out other funders | Is this AI safety area neglected? |
| Tractability | Can the team execute and can the field absorb the dollars? | Track record, institutional home, talent | No plausible execution path | Is the bottleneck funding, talent, or paradigms? |

| Criterion | Question | Evidence to seek | Red flags | Analogous AI safety question |
|---|---|---|---|---|
| Counterfactuality | What would happen without this grant? | Alternative funders, person's fallback career | Grant funds something that would happen anyway | Would the researcher join a lab or a safety org anyway? |
| Measurement | Can we observe progress and outcomes? | Career transitions, trial progression, detection uptime | Unmeasured or self-reported outcomes | Evaluation construct validity |
| Dual-use and downside risk | Could the work be misused or cause a release? | Containment, biosafety review, screening | No explicit dual-use assessment | Could the work accelerate misuse? |
| Political feasibility | If the output is policy, can it be adopted? | Coalitions, champions, stakeholder mapping | Technical soundness without feasibility | Can the governance proposal pass? |

ERA AIxBio Research Fellowship. P04 reports the program came inbound, had accepted AI and biosecurity fellows before, targeted a talent-constrained intersection, and was funded in slightly less than two months after assumptions were tested, track record was checked, the theory of change and downsides were examined, more information was requested, and input was given on structure. Rubric reading: strong on theory of change and track record; counterfactual is the fellows' alternative careers; measurement is career transitions and research outputs; dual-use risk is the main downside.

Cures Within Reach. P07 describes funding low-cost proof-of-concept trials for generic drug repurposing, a market failure because private pharma cannot establish exclusivity, and CG grantmakers funding LMIC-focused calls and follow-on trials. Rubric reading: strong on neglectedness and cost-effectiveness; measurement is trial progression and health outcomes; counterfactual is private pharma absent. But this is a health R&D grant, not a catastrophic-risk grant. The evaluator must ask whether it belongs in biosecurity or global health.

Strep A Vaccine Fund. S04 reports more than 600k deaths per year, about $14m per year of global R&D before the fund, about 50 times less than malaria with similar mortality, more than $140m raised at launch, and a goal to double clinical candidates and have at least one ready for Phase 3 by the end of 2030. Rubric reading: strong on problem size and neglectedness; tractability rests on

new human challenge models and earlier diagnostics; measurement is clinical candidates and Phase 3 readiness; counterfactual is other funders; dual-use is low but vaccine platform governance still matters.

```python
# BOTEC: compare $10m to AMF and a hypothetical bio x-risk grant
# Source anchor: S01 says $212m to AMF averts 67,000 deaths.
cost_per_death_amf = 212_000_000 / 67_000
donation = 10_000_000
lives_saved_amf = donation / cost_per_death_amf

# Hypothetical bio grant: reduces probability of a 1e9-death
    catastrophe by p
catastrophe_size = 1_000_000_000
for p in [1e-6, 1e-5, 1e-4, 1e-3]:
    expected_lives = p * catastrophe_size
    print(p, expected_lives)

# Break-even p: lives_saved_amf / catastrophe_size
break_even_p = lives_saved_amf / catastrophe_size
print(break_even_p)
```

*Nontrivial result from the BOTEC: at the S01 cost-per-death-averted, $10m to AMF saves about 3,160 lives. For a $10m bio grant against a 1e9-death catastrophe to match that, it must reduce the probability by about 3e-6. If the grant reduces probability by 1e-4, expected lives saved are about 100,000. The break-even is therefore a useful calibration, not a decision rule: catastrophe size, probability reduction, and correlation with other risks are all uncertain.*

*Caveat: the BOTEC is a proposed exercise, not a source-supported CG conclusion. The AMF figure is for a specific grant, not the marginal AMF dollar, and it omits non-x-risk benefits of biosecurity grants.*

## Sub-Question 4: Disputed Estimates and How to Represent Them

What are the disputed estimates in biosecurity, and how should an evaluator represent them without merging organizations into one voice?

### Step 1

Table 67: Table 4. Disputed estimates and positions. The table preserves the dispute rather than resolving it.

| **Claim** | **Source** | **Stated quantity or position** | **Object actually measured** | **Principal uncertainty** | **Who disputes it** |
|---|---|---|---|---|---|
| AI-bio uplift | P08 | One paper says LMs could help people with little or no lab training develop new pathogens; a RAND paper says current LMs did not substantially increase biological attack risk | Model capability uplift under different conditions | Scaffolding, model access, and real-world execution | The two funded research groups |
| Mirror bacteria risk | P06 | Andrew Snyder-Beattie's work on reducing risks from mirror bacteria | A novel synthetic-biology risk | Feasibility, containment, and timeline | Not specified in corpus |

| Claim | Source | Stated quantity or position | Object actually measured | Principal uncertainty | Who disputes it |
|---|---|---|---|---|---|
| IPCC-like biotechnology risk body | N14 | Failed because there was no consensus on the problem, incentives were misaligned, and no high-level experts would champion it | Political feasibility and expert consensus | Whether groundwork could create consensus later | Experts and institutions described in N14 |
| Strep A burden and neglect | S04 | More than 600k deaths per year; about $14m per year of R&D before the fund; about 50 times less than malaria | Mortality burden and R&D spend | Burden estimates and counterfactual R&D | Not specified in corpus |
| Generic drug repurposing market failure | P07 | Private pharma rarely pursues repurposing because it cannot establish exclusivity | Market incentive | Whether the drugs would otherwise be developed | Not specified in corpus |

*Nontrivial insight: not every dispute is a crux. A crux is a possible fact that is uncertain, decision-relevant, and resolvable at reasonable cost. The AI-bio uplift dispute is a crux if the grant decision hinges on it. The mirror-bacteria feasibility dispute may be a crux for a mirror-bacteria grant but not for a broad biohardening grant. The IPCC-like body dispute is a political-feasibility crux. Evaluators should fund evidence on the crux rather than take a side.*

Source-supported contrast: P08 funded both sides of the AI-bio disagreement. That is a portfolio approach to an unresolved crux. N14 shows that a technically sound policy can fail without political feasibility, so an evaluator should not treat expert consensus as a given.

## Sub-Question 5: Generalist Comparison Across Biosecurity, AI Safety, and Global Health

How should a generalist evaluator compare biosecurity with AI safety and global health without losing GCR focus?

### Step 1

Table 68: Table 5. Portfolio comparison. Figures are not normalized; the table is an exercise in structured comparison.

| **Dimension** | **Biosecurity** | **AI safety and security** | **Global health and wellbeing** | **Source anchor** |
|---|---|---|---|---|
| Theory of change | Prevent, detect, and respond to biological catastrophes; four-pillar resilience plan | Reduce worst-case AI risks through visibility, safeguards, and capacity | Evidence-backed health interventions | P01, P06, P08, S01 |
| Funding scale | More than $150m in 2026 | On track for over $1b in 2026 | Large and mature; S01 records a $212m single grant | P01, S01 |

| Dimension | Biosecurity | AI safety and security | Global health and wellbeing | Source anchor |
|---|---|---|---|---|
| Neglectedness | Very neglected in R&D; Strep A had about $14m per year before the fund | Young but growing rapidly; large funders entering | Large existing funder base and evidence base | S04, P01 |
| Tractability | Varies; detection and medical countermeasures more tractable than preventing deliberate misuse | Empirical safety work became more tractable as capabilities advanced | Highly tractable for bed nets and similar interventions | P06, S01 |
| Measurement | X-risk reduction is hard; career transitions and trial progression are easier | Evaluations and benchmarks are measurable, but construct validity is uncertain | Deaths averted and cost per life saved are measurable | P04, P07, S01 |
| Talent bottleneck | Dual expertise and lab access; field building is a major strategy | Staff time and competition with AI company salaries | Larger existing pipeline | P04, P06 |
| Risk of harm | Dual-use and biosafety risks | Misuse and loss of control | Low for bed nets; governance risks remain | P08, N14, P07 |

*Nontrivial insight: neglectedness is not importance, and tractability is not certainty. A generalist should avoid cause-prioritization paralysis by using the fund's own criteria and by treating the comparison as a portfolio-allocation question, not a single-winner contest. CG uses worldview diversification and*

*marginal returns; when every fund has more money, the emphasis shifts toward ambitious ownership and field-building. This is an inference from P06.*

- Biosecurity is a small, growing, talent-constrained portfolio where the four-pillar plan is a resilience bet.
- AI safety is larger and more technically diversified, with a stronger empirical-safety focus.
- Global health has the most mature measurement and the largest funding base.
- **A generalist evaluator adds value through reasoning transparency, calibration, and sourcing, not by pretending domain expertise.**

*Caveat: the comparison is a proposed exercise. The figures are sourced but not normalized, and the table should not be treated as a CG institutional ranking.*

## Sub-Question 6: Talent Bottlenecks and Proportionate Due Diligence: The ERA AIxBio Case

What are the biosecurity talent bottlenecks, and what does the ERA AIxBio case teach about proportionate due diligence?

### Step 1

Talent bottleneck: biosecurity needs dual expertise in biology plus security, policy, or governance; often lab access; sometimes clearances; and it faces a dual-use chilling effect. P04 describes field building and individual transition grants, and says there was just so much to be done. P01 says staff time is the main thing preventing CG from scaling. P06 says CG is hiring aggressively.

- The ERA AIxBio proposal arrived inbound rather than being solicited.
- The program had a track record with AI and biosecurity fellows.
- The grantmaker tested assumptions about the talent constraint.
- The grantmaker checked the track record with colleagues.
- The grantmaker thought through the theory of change and potential downsides.
- The grantmaker asked for more information and gave input on program structure.
- Slightly less than two months passed from proposal to grant.

*Nontrivial insight: proportionate due diligence scales with grant size, reversibility, and downside risk. A $50k individual transition grant can be justified by one conversation and a couple of reference checks, per P04. A multi-million-dollar program needs more. But biosecurity downside risk includes dual-use and*

*biosafety, so even a small grant may need checks on containment, institutional affiliation, and whether the work could be misused. The generalist's move is to ask what evidence would change the mind, what the counterfactual career is, and what the failure modes are.*

Source-supported contrast: P04's Trevor Levin habit is to ask early and often what piece of evidence would be most likely to change his mind, and to seek it out. That habit transfers directly to biosecurity, where the evaluator may not be able to judge the science but can judge the decision-relevance of the evidence.

*Caveat: the ERA case is a single inbound proposal. It is not a base rate for biosecurity grantmaking, and the corpus does not report the ERA grant size or its measured outcomes. Do not treat it as a proven success.*

# Build a global health, wellbeing and science R&D brief.

## Question

You are preparing for a generalist evaluator or grantmaker interview at Coefficient Giving. You must be able to place a bed-net campaign, a generic-drug repurposing programme and a global-aid policy think tank on a single cost-effectiveness logic, and then, with reasoning transparency, say exactly where each estimate could be wrong. Build a brief that does this for the global health, wellbeing and science R&D portfolio, at the level of precision a GCR-trained evaluator would demand from an AI-safety grant.

## Background

The global health and wellbeing portfolio is the older sibling of the GCR portfolio and supplies the cleanest worked examples of cost-effectiveness reasoning in the corpus. P06 reports that grants to evidence-backed global health interventions have saved over 100,000 lives and will save tens of thousands more, and that CG supported the late-stage clinical trials for the R21 malaria vaccine. The CEO is explicit about attribution: the credit largely belongs to program officer Katharine Collins, who co-invented the vaccine as a graduate student, rather than to the later funding. That sentence is itself a lesson in reasoning transparency.

The portfolio is not one intervention. P07 lists GiveWell's recommendations, Cures Within Reach and the Center for Global Development as individual staff suggestions, with the explicit disclaimer that they are not institutional positions and may not be the strongest option even within their own area. The same post records the operative comparability claim: GiveWell's marginal opportunity is expected to be roughly as cost-effective as CG's other marginal opportunities in global health.

The freshest worked example is S01: a $212 million grant to the Against Malaria Foundation for bed-net campaigns in the Democratic Republic of the Congo, with other GiveWell donors adding $64 million, and an expectation of 67,000 deaths averted. The same post cites two WHO population aggregates, 14 million malaria deaths averted since 2000 and a 70 percent drop in African malaria cases between 2000 and 2015 that WHO credits substantially to net use, and notes that nets cost a few dollars each with a couple more for distribution.

The reference frameworks for grading any of these claims are D01, the reasoning-transparency standard, and D02, which presents the back-of-the-envelope calculation (BOTEC) as a calibration tool rather than a verdict. Together they supply the standard against which an AMF grant review and the staff suggestion list must be assessed.

## Sub-Question 1: Cost-Effectiveness Logic and What Comparable Actually Claims

Before any grant can be ranked, the ratio must be fixed. A global health grant is not sold as good; it is sold as a number of dollars per unit of impact, against an explicit baseline, under stated uncertainty. The exercise is a ratio with three parts, and each part has a different failure mode.

### Step 1

The arithmetic is trivial; the definitions are not. Write the ratio explicitly, because almost every disagreement about a grant is really a disagreement about one of its three parts.

$$\text{CEA} = \frac{\text{impact} - \text{impact under the counterfactual baseline}}{\text{grant dollars} + \text{cost imposed on others}}$$

Each of the three parts is a different object, and each is checked by a different piece of due diligence.

- **Numerator:** the hazard reduction attributable to the intervention, converted into a final unit such as deaths averted, DALYs, or CG dollars as P03 uses.
- **Baseline:** the state of the world absent this grant, including nets that other actors would have funded anyway.
- **Denominator:** the grant plus the costs the project imposes on others. D04 lists this explicitly as any additional costs the project may impose on others, such as the time required from evaluators or collaborators.

The units are not interchangeable. Global health reports deaths averted; P03 reports welfare in CG dollars, defining a one percent income increase for one person for one year as CG$500. Moving between the two requires a moral weight, and the choice of weight is a view, not a measurement. This is the first non-trivial jump: a ratio without a baseline is not a cost-effectiveness claim, it is an impact report.

GiveWell's AMF review makes the baseline structural. As D01 records, the review asks whether LLINs are targeted at people who do not already have them and whether there is room for more funding. *Both are counterfactual questions wearing the costume of due diligence*, and they are the reason GiveWell's number survives comparison with a naive deaths-per-dollar calculation.

**Step 2**

P07's comparability claim is precise and narrow: GiveWell's marginal opportunity is expected to be roughly as cost-effective as Coefficient's other marginal opportunities in global health. That is a within-cause claim. P06 makes a broader and softer claim: each additional dollar should produce roughly the same amount of good if allocated to any of our funds.

These are not the same claim, and the difference matters at interview. A candidate who repeats the slogan as if it were a measured equality will be asked to defend a cross-worldview ranking they have not built.

Table 69: *Read the source claim before repeating the slogan; the three rows are nested, not equivalent.*

| **Scope** | **Source claim** | **What it licenses** | **What it does not license** |
|---|---|---|---|
| Within global health | P07: GiveWell's marginal opportunity roughly as cost-effective as CG's other marginal global-health opportunities | Treating GiveWell as a live benchmark for reallocation inside the fund | Any ranking against GCR or AI-safety grants |
| Across funds | P06: each additional dollar should produce roughly the same good if allocated to any fund | An aspiration that holds under a single moral weight | A measured equality; P06 keeps worldview diversification explicitly alongside it |
| Across worldviews | P06 and P08: staff disagree with each other and CG supports actors who disagree with CG | Multiple defensible orderings of the same dollars | A single cardinal ranking of lives saved against catastrophic-risk reduction |

The non-trivial consequence: cost-effective is undefined until a moral weight is chosen, and the corpus says openly that reasonable staff disagree about the weight. An answer that ranks bed nets against AI safety without naming the weight is making a hidden value judgement dressed as a calculation.

## Sub-Question 2: From Grant to Expected Deaths: the AMF / DRC Case

S01 gives one grant and three numbers: $212 million from CG, $64 million from other GiveWell donors, and 67,000 deaths averted. Two of those are commitments and the third is a forecast. Do the arithmetic first, then interrogate the forecast and the population aggregates that surround it.

### Step 1

The grant announcement bundles two funders. The correct denominator for the CG decision is the CG commitment; the correct denominator for the intervention's cost-effectiveness is the combined commitment, because the co-funding is applied to the same campaign. Keep the two apart in the note and cite the combined figure only when computing the intervention's ratio.

$$\text{total} = 2.12 \times 10^8 + 6.4 \times 10^7 = 2.76 \times 10^8 \qquad \frac{2.76 \times 10^8}{67{,}000} \approx \$4{,}119 \text{ per death averted}$$

Notice what this figure is and is not. It is a mean over a modeled distribution, not an observed count, and it is roughly linear in the mortality parameter. S01 does not publish an interval, so the honest move is to show the ratio's sensitivity rather than to quote a single point.

```
cg_grant = input_state.get("cg_grant", 212_000_000)
co_funding = input_state.get("co_funding", 64_000_000)
deaths_averted = input_state.get("deaths_averted", 67_000)

total_committed = cg_grant + co_funding
cost_per_death_averted = total_committed / deaths_averted

deaths_grid = [40_000, 50_000, 67_000, 80_000, 100_000]
sensitivity = {d: total_committed / d for d in deaths_grid}

output_state["total_committed"] = total_committed
output_state["cost_per_death_averted"] = cost_per_death_averted
output_state["cost_per_death_sensitivity"] = sensitivity
```

The mean lands in the same order of magnitude as GiveWell's other headline ratios, which is the whole point: the AMF grant is not interesting because $4,119 is a surprising number, it is interesting because the number is defensible at that order of magnitude across a decade of replication.

### Step 2

S01 adds two population aggregates: 14 million malaria deaths averted since 2000, and a 70 percent drop in African malaria cases between 2000 and 2015

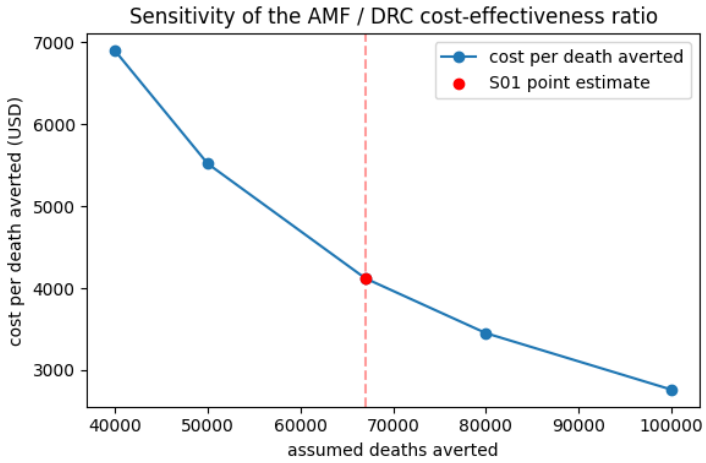


Figure 37: Graphical analysis for Step 1.

that WHO credits substantially to net use. Both are ecological associations, not experiments on the marginal net.

Two inferential gaps separate them from the effect of the next net funded.

- **Aggregation across actors.** Nets, artemisinin combination therapy, indoor residual spraying, diagnostics and income growth all moved together over the period. Attributing the trend to nets requires an identification strategy, not a co-movement.
- **Aggregation across time.** The 2000 to 2015 decline is a period average. Insecticide resistance and residual transmission erode the yield of a marginal net today, so the historical average is a ceiling on the current marginal effect, not an estimate of it.

This is the same error we flagged in the threat-model work: an association that is valid in one row of a matrix is being used to license an estimate in a different row. There the row was deliberate-misuse evidence; here the row is the historical population trend, and the target is the marginal grant.

One more sanity check is available from S01 itself. At a few dollars per net plus a couple more for distribution, the total implies tens of millions of nets, and the deaths-averted figure implies a plausible nets-per-death range consistent with the wider literature. If the arithmetic did not land in that range, the ratio would be suspect before any causal question was asked.

```
total_committed = input_state.get("total_committed", 276_000_000)
deaths_averted = input_state.get("deaths_averted", 67_000)
cost_per_net_low = input_state.get("cost_per_net_low", 5.0)
cost_per_net_high = input_state.get("cost_per_net_high", 8.0)
```

```
5
6  nets_low = total_committed / cost_per_net_high
7  nets_high = total_committed / cost_per_net_low
8  nets_per_death_low = nets_low / deaths_averted
9  nets_per_death_high = nets_high / deaths_averted
10
11 output_state["nets_distributed_range"] = (nets_low, nets_high)
12 output_state["nets_per_death_averted_range"] = (nets_per_death_low,
       nets_per_death_high)
```

*The inference for an interview: use the WHO figures for direction, the AMF model for the magnitude, and never the WHO figures to validate the AMF forecast. The two answer different questions.*

## Sub-Question 3: Market-Failure Justifications: Which Failure, and Is It Binding?

Each of the three P07 suggestions rests on a market failure, and they are not the same failure. Naming the correct one determines which evidence is relevant and which is a distraction. Then ask whether the failure is binding on your dollar specifically.

### Step 1

Classify before justifying. The AMF bed-net case, the Cures Within Reach drug-repurposing case and the Center for Global Development policy case each rely on a different failure of the private market.

Table 70: Three distinct failures licensing three distinct grants; calling them all market failures hides which remedy is appropriate.

| **Instrument** | **Grant (source)** | **Market failure** | **Mechanism** | **Why philanthropy is the residual buyer** |
|---|---|---|---|---|
| Public health good with an externality | AMF bed nets (S01) | Positive externality plus coordination failure | Nets protect non-users by reducing transmission, and households do not price that in | Retail does not deliver the herd-level subsidy; a large coordinated buyer must step in |

| Instrument | Grant (source) | Market failure | Mechanism | Why philanthropy is the residual buyer |
|---|---|---|---|---|
| Non-appropriable drug | Cures Within Reach (P07) | Missing market due to non-excludability | Generic drugs cannot be patented for new indications, so no firm can recoup the trial cost | Philanthropy is the buyer of last resort for a good the pharmaceutical market structurally will not produce |
| Public-good policy research | Center for Global Development (P07) | Non-excludable information good | Evidence-based aid policy is non-rival and cheap to copy once produced | No individual beneficiary will pay for research that benefits a diffuse set of future governments and donors |

P08 gives the institutional framing for why a GCR-focused funder would recognise these as legitimate adjacent work: there are sometimes market failures around the upsides, particularly around accelerating global public goods that lack large markets, such as scientific research and products that address neglected diseases. The three suggestions are instances of exactly that framing applied outside the AI portfolio.

S04 uses the same three-part read-out explicitly for the Strep A Vaccine Fund: a huge problem, at roughly one percent of total annual deaths worldwide; super neglected, at about $14m per year of global R&D before the fund, roughly 50 times less than malaria which has similar mortality; and newly tractable, via new human-challenge models and earlier diagnostics. Neglect and tractability are the observable consequences of a market failure.

**Step 2**

A market failure explains why the good exists in insufficient quantity. It does not establish that a particular dollar adds anything. Additionality is a separate quantity and must be estimated separately, or the cost-effectiveness ratio is measuring the wrong thing.

$$\mathrm{EV}_{\text{grant}} = \Pr(\text{additionality}) \times \frac{\text{impact of the marginal unit}}{\text{dollars}}$$

D01 records the four counterfactual questions in GiveWell's AMF review: are LLINs targeted at people who do not already have them; do LLINs reach intended destinations; is there room for more funding; and how generally effective is AMF as an organization. The first three are additionality and fidelity checks, not mechanism checks.

GiveWell concedes displacement directly: it is likely that some portion of AMF's funding has displaced other funding into other malaria interventions and into other uses. That is not fatal. Displaced dollars may be nearly as good, and GiveWell's own treatment of the question is a model of reasoning transparency. But it contaminates the naive ratio, because the ratio assumes the net would not have been distributed otherwise.

```
base_cpe = input_state.get("cost_per_death_averted", 4_119.0)
additionality = [0.5, 0.7, 0.8, 0.9, 1.0]

adjusted = {a: base_cpe / a for a in additionality}

output_state["cost_per_death_adjusted_for_additionality"] = adjusted
```

*At fifty percent additionality the true cost per death averted doubles; at one hundred percent it matches the headline. The general rule this inoculates against, and the answer a GCR interviewer will expect, is: funding a good thing without checking whether someone else would have funded it is not yet a grant recommendation.*

## Sub-Question 4: Measurement Caveats: Attribution, Displacement and Counterfactuality Are Three Different Errors

Attribution, displacement and counterfactuality are routinely merged in one sentence in popular summaries, and they are three different errors with different signs, different magnitudes and different remedies. Separate them before saying anything that sounds like a caveat.

### Step 1

Each error asks a different question and is fixed by a different study design or due-diligence step.

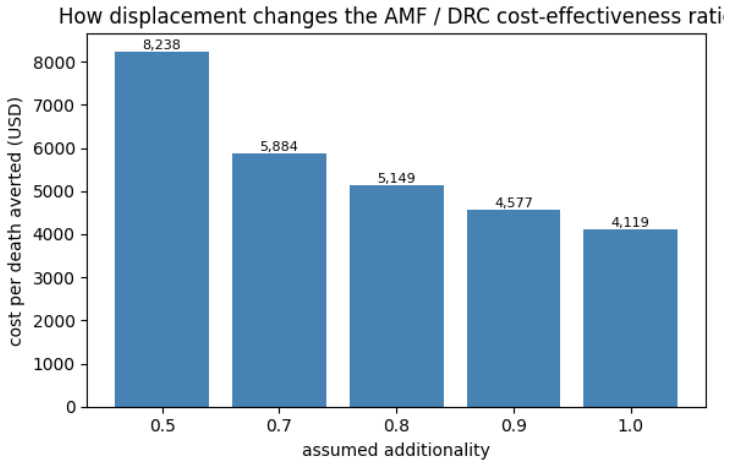


Figure 38: Graphical analysis for Step 2.

Table 71: Conflating these three is the commonest failure mode in a candidate who is trying to sound critical of a headline number.

| **Error** | **The question** | **Corpus anchor** | **What would move the estimate** |
|---|---|---|---|
| Attribution | Did our funding cause the outcome? | P06's over-100,000-lives figure; S01's claim that WHO credits net use for the 70 percent decline | A controlled comparison with a no-funding arm, or a track record with a stated counterfactual |
| Displacement | Would someone else have funded this? | GiveWell's own AMF review concedes some funding has displaced other malaria funding and other uses | Other funders' response functions; coverage at the margin; room for more funding |
| Counterfactual baseline | Deaths averted relative to what? | Every dollars-per-death-averted figure in S01 and P06 | Baseline transmission, bed-net coverage and insecticide resistance in the target districts |

*A useful self-test: if your caveat could be restated as the same sentence with the word attribution replaced by displacement and it still sounds correct, you probably have not distinguished them.*

### Step 2

Apply D01's reasoning-transparency standard to the three numbers in play. Each deserves a different level of confidence, and a candidate who collapses them into one number is failing the standard.

- **Over 100,000 lives saved (P06)** is a track-record aggregate over many interventions and more than a decade. It is a strong directional claim about the portfolio; it is not a cost-effectiveness estimate and should not be quoted as one.
- **67,000 deaths averted (S01)** is a forecast from an explicit, inspectable model. It is the most defensible of the three for a marginal decision because its assumptions are named, even though S01 does not publish an interval.
- **14 million deaths averted since 2000 and the 70 percent drop (S01)** are population aggregates. They are the weakest of the three for causal purposes and the strongest for direction, and they should be used only to establish that the intervention class works at all.

What would move the marginal estimate, stated in advance, in the D01 style.

- **Upward:** a randomized comparison showing nets still reduce all-cause child mortality at scale under current resistance conditions, or an additionality design showing the marginal net reaches previously uncovered households.
- **Downward:** evidence that the marginal net goes to households already covered, or that co-funders defund malaria as AMF expands.

```
import matplotlib.pyplot as plt

base = input_state.get("cost_per_death_averted", 4_119.0)

factors = {
    "mortality estimate (-30% to +30%)": (base * 0.7, base * 1.3),
    "additionality (1.0 to 0.5)": (base, base / 0.5),
    "net cost (-40% to +40%)": (base * 0.6, base * 1.4),
}

names = list(factors.keys())
lows = [factors[n][0] for n in names]
highs = [factors[n][1] for n in names]

fig, ax = plt.subplots(figsize=(7, 4))
for i, n in enumerate(names):
    ax.barh(i, highs[i] - lows[i], left=lows[i], height=0.5)
```

```
ax.axvline(base, color="black", linestyle="--", label="point
    estimate")
ax.set_yticks(range(len(names)))
ax.set_yticklabels(names)
ax.set_xlabel("cost per death averted (USD)")
ax.set_title("Tornado view of what moves the AMF / DRC ratio")
ax.legend()
plt.tight_layout()
plt.savefig(_generated_image_path_)
```

A calibrated interview script that survives the D01 standard: Coefficient's evidence-backed global health grants are credited with over 100,000 lives saved to date, and the AMF grant for DRC bed-net campaigns is expected to add roughly 67,000 more at about $4,100 per death averted on the combined $276 million commitment. I would weight the historical track record more heavily than the forecast, and the forecast more heavily than the WHO population aggregates, because the forecast has a named model and a defined counterfactual while the aggregates do not. My main residual uncertainty is additionality: if other funders would have paid for the same nets, the true cost per death is higher, and I do not know the displacement rate. I would treat that uncertainty as the thing to resolve before renewing at scale rather than as a reason to withhold funding now.

Note the deliberate phrasing: *is credited with, is expected to, on the combined commitment, I would weight, my main residual uncertainty.* Each phrase signals where the assertion stops and the estimate begins, which is exactly what D01 asks for.

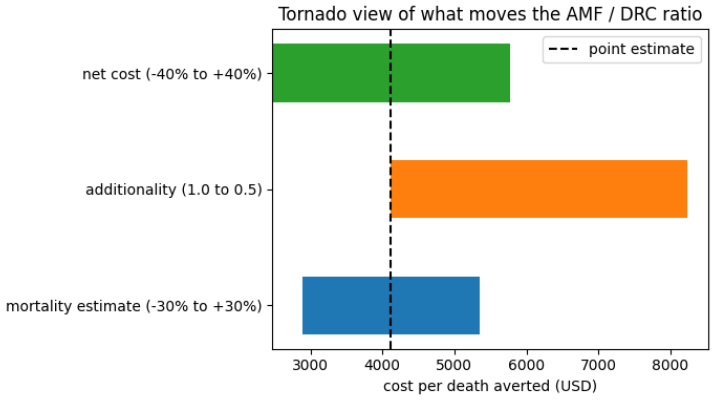


Figure 39: Graphical analysis for Step 2.

# Brief on farm animal welfare and Effective Giving & Careers.

## Question

Prepare a portfolio-breadth brief covering two adjacent Coefficient Giving areas: **farm animal welfare** and the **Effective Giving & Careers fund** (which contains an effective-giving side and a careers or talent side). For each area the brief must give the theory of change, the metrics and the counting conventions behind the headline numbers, the instruments and grantees, and the counterfactuality, attribution and persistence problems that a generalist or Global Catastrophic Risks interviewer would expect you to name unprompted.

Deliverable standard: every claim must be labelled **source-supported**, **inference**, or **proposed exercise**. Individual staff suggestions must not be presented as institutional positions, and single-case evidence must not be presented as a base rate.

## Background

This is phase 4 of 6, step 3 of 6. The phase purpose is portfolio breadth: because CG is multi-fund and an interviewer may test generalist judgment, you need the non-AI portfolios well enough to compare opportunities, assess cost-effectiveness, and reason across cause areas without losing GCR focus.

Primary source anchors for this step: **P06 (2025 CEO letter)** for farm animal welfare history and scale; **P05 (Effective Giving & Careers 2026 strategy)**, **P02 (Effective Careers RFP reflection)**, and **P07 (individual donor suggestions)**. Secondary anchors: P04 (day in the life, for how grant investigations actually run), S10 (Austin Chen on philanthropic funding pitfalls), and P03 (a CG welfare-unit convention used elsewhere in the corpus).

Discipline note carried throughout: P07 states that its suggestions come from individual program staff or teams, do not necessarily represent Coefficient's institutional view, and are not claims that the suggested option is the best across areas or even within its own area. Treat P07 as a set of attributed judgments with named recommenders, useful for inferring what the field considers strong, not as a fund strategy document.

## Sub-Question 1: Farm Animal Welfare: Chain of Leverage, Counting Unit, and the Persistence Conditional

P06 reports two numbers. Past: 'Our work so far has improved the lives of over 3 billion farm animals, largely by driving commitments from corporations to improve conditions in their supply chains.' Future: 'if the reforms they've helped secure stay in place (even putting aside future wins), we project that they will improve the lives of another 10 billion animals over the next 10 years.'

*Note what carries the weight in the second sentence. The projection is conditional, and the condition ('if the reforms stay in place') is a statement about a governance and enforcement variable that is not quantified anywhere in the corpus. The arithmetic is not the claim; the condition is the claim.*

P07 recommends the EA Animal Welfare Fund, attributed to the Farm Animal Welfare team, on three grounds: it funds the most promising neglected opportunities, it has an open application process that serves individuals and new groups regardless of networks, and its managers are considered among the most experienced and thoughtful researchers and advocates in the movement. Note that this is a team attribution, so it is closer to institutional than a single-staff pick, but it would still be an overread to call it fund strategy.

P06 also supplies the field-building and personnel-judgment point: Lewis Bollard was hired in 2015 to lead farm animal welfare giving, and CG wrote publicly at the time about the worry that, a couple of years out of law school, he might not have enough experience to succeed in the role. This is a clean example of an early bet whose value could not be assessed at the time of the grant, which is the same structure as the bet on a small careers or effective-giving organization.

**Step 1**

**Step 1: state the causal chain and mark where the corpus stops.**

Table 72: The chain has three links and the corpus quantifies the first and the third, not the second.

| **Link in the chain** | **What the corpus gives us** | **Status** |
|---|---|---|
| Corporate commitment obtained | P06: driving commitments from corporations to improve conditions in their supply chains is the stated mechanism | Source-supported mechanism |
| Commitment implemented and held | P06 speaks of reforms staying in place and decoupling the projection from future wins; no implementation or reversion rate is given | The operative assumption |
| Animal lives improved | P06: over 3 billion so far; a projected 10 billion more over 10 years conditional on persistence | Source-supported conditional projection |

**Step 2: the counting unit. The headline is in animals, not in any welfare unit that is commensurable with the units used elsewhere in the same corpus.**

Table 73: The corpus supplies a human welfare numeraire in one area and a headcount in another, and never converts between them.

| Counting convention | What it counts | Why it matters for comparison |
|---|---|---|
| Animals affected | Headcount of animals whose conditions changed | Counts bodies, not time; a broiler chicken lives roughly six weeks, so headcount converts into few animal-years |
| Animal-years of improved conditions | Headcount times the duration of the improvement | The convention that makes species comparable to one another; not published in this corpus |
| CG dollar welfare units | P03's convention for human income: a 1 percent income increase for one person for one year is worth CG$500 | No published animal-to-CG-dollar conversion appears in this corpus, so cross-area ranking is blocked at the currency step |

**Step 3: make the conditional explicit rather than leaving it buried in prose.**

```
# Read source-stated quantities from shared state.
animals_projected_10y = input_state.get("animals_projected_10y",
    10_000_000_000)
animals_to_date = input_state.get("animals_to_date", 3_000_000_000)

# P06's projection is conditional on reforms staying in place. Treat
    that
# condition as an explicit free parameter rather than a hidden
    assumption.
persistence_grid = [0.25, 0.5, 0.75, 1.0]
conversion = input_state.get("implementation_conversion", 1.0)  #
    share of commitments that bite

faw_grid = {p: animals_projected_10y * p * conversion for p in
    persistence_grid}

# By construction the headline equals the fully-persistent case.
```

```
faw_headline_if_persistent = animals_projected_10y * conversion

output_state["faw_persistence_grid"] = {
    "persistence_{}".format(int(p * 100)): v for p, v in
    faw_grid.items()
}
output_state["faw_headline_if_persistent"] =
    faw_headline_if_persistent
output_state["faw_animals_to_date"] = animals_to_date
```

The arithmetic is trivial and that is exactly the point. The output is a grid, not an estimate: what a grantmaker can buy is evidence about the persistence parameter, not about the multiplication. If asked 'what would you want to know before funding more of this', the correct answer is a monitoring claim, not a magnitude claim.

**Step 4: proposed exercise. Rank these checks by how much they would move the persistence parameter, and say what evidence type each requires.**

- Are the commitments contractually binding with dated milestones? Evidence type: primary documents, not summaries.
- Do independently audited cage-free or crate-free shares keep moving after the announcement year? Evidence type: repeated third-party audits over several years, which is the only design that can detect reversion.
- Does the commitment cover the whole supply chain or only a subset of purchases? Evidence type: procurement scope documents; this is where a large headcount can shrink without any reversion.
- What happens when the public campaign pressure ends? Evidence type: a comparison of firms whose commitments were announced under pressure with firms that adopted similar standards without pressure, which is the counterfactual question reused from Step 5 below.

### Step 2

Step 5: the counterfactuality of a commitment. A corporate commitment to improve supply-chain conditions is only additional if the firm would not have adopted the standard anyway, in the same window. Note the structural echo of the prior phase: this is the same question as whether a grant moved money that would have moved anyway, and it has the same answer shape, which is that you need a comparison group and not a before-and-after.

**Two defensible answers to 'should CG discount animal-welfare headcounts for counterfactuality when it reports progress?'**

- Answer A: no discount in the headline, because the headline is a description of animals affected, not a causal claim, and mixing a causal adjust-

ment into a descriptive count destroys the ability to compare the description across years. This answer depends on the reporting purpose being communication of scale.

- Answer B: discount, because the number is used inside a cost-effectiveness comparison and P05 explicitly applies counterfactuality adjustments when computing portfolio returns on the effective-giving side, so a comparable adjustment is required for comparability across funds. This answer depends on the number being used for allocation rather than for communication.

The two answers are reconcilable in the way the corpus actually behaves: publish the descriptive count for scale, and state the causal adjustment separately for allocation. That reconciliation is an inference built on P05's disclosure practice, not a documented CG policy.

## Sub-Question 2: Effective Giving: Which Return on Investment, and Why the Same Word Means Two Different Things

P05 describes the effective giving side: organizations supported 'have collectively moved hundreds of millions of dollars to cost-effective charities'; the 2025 effective giving RFP funded 10 organizations that 'collectively moved ~$14m'; 'After incorporating quality and counterfactuality adjustments, we obtained a weighted average ROI of ~5x for the portfolio'; and 'our full effective giving portfolio is strong, having moved ~$300m in 2025, at an average weighted ROI of ~5-6x.'

P02, describing the adjacent careers RFP, reports that 'Weighting each grant's modeled social return on investment by its funding amount gives a mean of roughly 5,500x - though these are early, uncertain estimates.'

Two numbers, one phrase, and a gap of three orders of magnitude. The nontrivial move is to see that this is not an error or a disagreement. The two figures are **different registers**: one is a leverage or flow ratio over money moved, the other is a modeled social benefit per grant dollar. Quoting either without its register is the single easiest way to fail a generalist interview.

**Step 1**

**Step 1: separate the registers before comparing anything.**

Table 74: The corpus contains three different denominations under one vocabulary.

| Claim as stated | Source | Numerator | Denominator | What it cannot support |
|---|---|---|---|---|
| weighted average ROI of ~5x for the portfolio; ~5-6x at portfolio level | P05 | Dollars moved by grantees to cost-effective charities, after quality and counter-factuality adjustments | Not stated in the source; the ambiguity between grant dollars and grantee operating cost is the crux | A claim about social welfare per dollar, because the numerator is a financial flow, not a welfare quantity |
| mean of roughly 5,500x, weighted by funding amount | P02 | Modeled social return of each grant | Funding amount of the grant | Comparability with the ~5x figure, since dollars of social benefit are not dollars moved; also covers only about 85 percent of recom-mended grants |

| Claim as stated | Source | Numerator | Denominator | What it cannot support |
|---|---|---|---|---|
| a 1 percent income increase for one person for one year is worth CG$500 | P03 | Human welfare in a common unit | One person-year at one percent income | It is a human-income convention; it supplies no animal conversion, so it cannot adjudicate the farm animal welfare comparison from the previous subquestion |

**Step 2: test the leverage reading against the program's own cumulative figures.**

```
money_moved_2025 = input_state.get("money_moved_2025", 300_000_000)  # P05
roi_low, roi_high = input_state.get("roi_range", (5.0, 6.0))  # P05
cumulative_spend = input_state.get("cumulative_program_spend", 30_000_000)  # P05, since 2022

# If ROI were a pure leverage ratio (dollars moved per grant dollar spent),
# the implied 2025 grant spend would be:
implied_low = money_moved_2025 / roi_high
implied_high = money_moved_2025 / roi_low

output_state["implied_grant_spend_2025_under_leverage_reading"] = (
    round(implied_low), round(implied_high)
)
output_state["cumulative_program_spend_since_2022"] = cumulative_spend
output_state["implied_to_cumulative_ratio"] = (
    round(implied_low / cumulative_spend, 2), round(implied_high / cumulative_spend, 2)
)
```

The leverage reading implies roughly fifty to sixty million dollars of grant spend in 2025 alone, against a reported cumulative program spend of more than thirty

million dollars across more than sixty grants since 2022. That is arithmetically inconsistent, so the leverage reading of the ~5x figure cannot be the whole reading.

*At least one of the following must therefore hold, and naming them is the answer to the interview question.*

1. The ratio is modeled social benefit per grant dollar with a welfare conversion, not dollars moved per dollar granted.
2. The denominator is the grantee's own fundraising or operating cost, not CG's grant, so the ratio describes the grantee's cost-effectiveness rather than the funder's leverage.
3. The three hundred million dollars is a cumulative or multi-year flow being described alongside a single-year ratio, so numerator and denominator cover different periods.
4. The two ROI figures cover different populations: the ten-organization RFP cohort versus the full portfolio including multi-year grantees funded before 2025.

The corpus does not let us choose among these. That is the honest answer, and it is stronger than picking one. The corpus does, however, tell us something about which reading P05 intends at the portfolio level: P05 applies quality and counterfactuality adjustments before reporting the ratio, and counterfactuality adjustments are a property of a model of causal contribution, not of a cash transfer. That leans against reading one, and it is an inference rather than a source statement.

**Step 3: two defensible answers to 'is the effective giving portfolio's ~5x a good number?'**

- Answer A: uninterpretable as stated, and the right move is to ask what the denominator is before comparing it to anything. This answer depends on treating the ratio as a decision input, where a three-order-of-magnitude ambiguity dominates.
- Answer B: good enough to act on, because the bar test in P05 is stated as 'well above our bar' and the same bar is used to decide renewals, so the number's function is ranking within a portfolio rather than absolute calibration. This answer depends on the comparison set being the fund's own alternatives rather than other cause areas.

## Sub-Question 3: Effective Careers: The RFP as Funder Instrument, Outputs Versus Outcomes, and Selection on Measurability

P02 is a reflection on the first open call for the effective careers side of CG's work: it awarded $13,075,301 across 28 organizations through the Effective Giving &

Careers Fund, with several other CG funds expecting to fund up to 16 more; it drew 474 applications, of which 111 advanced to a second stage; the funder had estimated a 25 percent chance of receiving more than 40 applications; weighting each grant's modeled social return by its funding amount gives a mean of roughly 5,500x on early uncertain estimates; 13 of the 28 grants are cause-neutral and 15 overlap strongly with one or two specific cause areas, with global health and development the heaviest concentration; and the funder found that many of the strongest opportunities came from impact-focused organizations outside the EA network.

P05 adds the strategy layer: more than 60 grants and more than thirty million dollars since 2022 across effective giving, talent, incubation and research; 2026 funding nearly three times prior years; Ambitious Impact launched eight new global health and development charities in 2025; new policy and research talent niches; and a plan to run two to four focused RFPs in 2026.

P07 adds concrete career metrics for one grantee: High Impact Professionals supported 210 accelerator participants and 519 users of a self-guided track, reported 51 career transitions and 11 fellowship transitions, and 67 participants taking a giving pledge with 18 committing at least ten percent of income.

**Step 1**

**Step 1: an RFP is not only a grantmaking channel. Treat it as an instrument with several outputs, only one of which is grants.**

- Dealflow: surfaces organizations the funder would not have known to solicit, which P02 states directly in the finding that many of the strongest opportunities came from outside the EA network.
- A landscape map: P05 states that the 2025 effective giving RFP helped map the landscape systematically and compare opportunities side by side.
- Prior correction: the 474-against-40 surprise is evidence that the funder's model of the field's size was wrong, which is an output in its own right and should be recorded as such in a portfolio review.
- Co-funding attraction and bar calibration, both stated in P05 for the effective giving RFP; P02 states that the common grants were evaluated on career transitions per unit of budget and measured against the fund's bar.
- A renewal signal: P05 says the RFP helped determine which grants warranted renewable support.

The non-trivial consequence: a large application volume is not the constraint the RFP tests. P01 states that staff time is the main thing preventing CG from scaling impact, and P02's own surprise implies that application supply is abundant. So the binding constraint is evaluator capacity, and an RFP that

raises application volume without raising screening capacity makes the bottleneck worse, not better.

**Step 2: separate outputs from outcomes, and attach the counterfactuality question to each.**

Table 75: P02 notes 'a wide range of how applicants monitor their outcomes' and that an internal project on shared standards is under consideration; the table is a proposal for what such standards would have to distinguish.

| Metric | What it actually measures | Counterfactuality risk | Evidence that would settle it |
|---|---|---|---|
| Applications received | Field size and funder salience, mildly contaminated by the funder's own marketing push | Low, since the number describes the applicant pool rather than an effect | Compare application counts across RFPs with and without active promotion, as P02 already plans for a future call |
| Career transitions | An output: a person changed role | High: the person may have moved anyway, and the receiving organization may count the same hire | A comparison group of comparable professionals who did not use the program, plus duplicate-claim reconciliation across grantees |
| Career plan changes | An intermediate: intent or stated plan | Very high: stated intent has a weak and selective relationship to subsequent action | Panel follow-up at twelve and twenty-four months, with a pre-registered threshold for what counts as a change |

**Step 3: the weighted mean is dominated by its largest grants. P02's figure is explicitly a funding-weighted mean, which is not the same as an unweighted mean across grants, and the difference grows with the spread of grant sizes.**

```
import statistics

# Illustrative, invented numbers: the arithmetic only, not a claim about the RFP.
grants = input_state.get("illustrative_grant_rois", [2.0, 3.0, 4.0, 6.0, 9.0, 25.0])
sizes = input_state.get("illustrative_grant_sizes", [2e5, 3e5, 4e5, 6e5, 1e6, 4e6])
assert len(grants) == len(sizes)

weighted = sum(g * s for g, s in zip(grants, sizes)) / sum(sizes)
unweighted = statistics.fmean(grants)

output_state["funding_weighted_mean"] = round(weighted, 2)
output_state["unweighted_mean"] = round(unweighted, 2)
output_state["divergence_ratio"] = round(weighted / unweighted, 2)

# Share of the weighted mean contributed by the single largest grant.
largest = max(zip(grants, sizes), key=lambda t: t[1])
output_state["largest_grant_share_of_weighted_mean"] = round(
    (largest[0] * largest[1]) / sum(g * s for g, s in zip(grants, sizes)), 3
)
```

The point is not the number; it is that a funding-weighted mean of ratios is a concentration statistic. Quoting it as 'the portfolio's return' silently answers a question about the average grant with a figure that is mostly about one grant. In an interview, say which object you are describing.

**Step 4: the two failure modes a generalist should name unprompted.**

Selection on measurability. Organizations with better monitoring capacity look better to an evaluator than organizations with identical effects and weaker reporting. P02 records wide variation in how applicants monitor outcomes, and P05 proposes funding organizations to strengthen monitoring and evaluation practices. That remedy reduces the reporting gap, but the portfolio's measured return is still computed on the subset that can be measured, which P02 flags by noting the estimate covers about 85 percent of recommended grants.

Double counting and credit allocation. S10 lists exactly this as a pitfall of philanthropic funding: unclear credit allocation and double-counting of impact. In a careers portfolio, the same transition can be claimed by a coach, a fellowship, a job board, and the hiring organization. The corpus does not describe how CG reconciles duplicate claims, so the honest statement is that this is an open question the portfolio needs, not a solved one.

P05 states the underlying tension in its own words among its key uncertainties: whether the program is 'appropriately balancing the shorter feedback loops of scaling donation flows against the longer-term, harder-to-measure impacts

of shaping careers, or are we being drawn too much toward what's easier to evaluate.' That is the funder conceding the selection problem. Use it as the source-anchored version of the inference.

## Sub-Question 4: Cross-Cause Comparison: Counterfactuality, Persistence, Credit Allocation, and the Missing Common Currency

This is the exercise the tiny step asks for: assess measurement pitfalls in career-transition and animal-welfare return-on-investment estimates, including counterfactuality, attribution, and the persistence of reforms, and then say what follows for a generalist evaluator who must hold both areas at once.

The structural point to carry into the interview: the same four pitfalls appear in both areas, but the areas differ in which pitfall is load-bearing. In farm animal welfare, persistence is load-bearing because the projection is explicitly conditional on it. In effective careers, counterfactuality and credit allocation are load-bearing because the unit of success is a person's decision, which is not owned by any single grantee.

**Step 1**

**Step 1: the four pitfalls, instantiated in both areas.**

Table 76: The same four pitfalls, with different load-bearing members in each area.

| **Pitfall** | **The question it raises** | **Farm animal welfare instance** | **Effective careers instance** |
|---|---|---|---|
| Counterfactuality | Would this have happened without the grant? | A firm may have adopted a supply-chain standard anyway, under consumer pressure or internal policy cycles | A professional may have moved into high-impact work without the coaching, fellowship, or job board |

| Pitfall | The question it raises | Farm animal welfare instance | Effective careers instance |
|---|---|---|---|
| Persistence | Does the measured change survive the end of the intervention? | The load-bearing case: P06's ten-billion projection is explicitly conditional on reforms staying in place | A person can leave the role, or the role can be reorganized, within a year of the transition |
| Attribution and credit | Who can claim the effect, and can the same effect be claimed twice? | A campaign, a funder, and a corporate sustainability team all plausibly claim the same commitment | A coach, a fellowship, a job board, and the hiring organization all plausibly claim the same hire; S10 names this as a known pitfall |
| Selection on measurability | Are we measuring the effect, or the ease of measuring it? | Corporations with audited supply chains are more countable than dispersed producers | Organizations that track alumni are more countable than those that do not; P02 reports a wide range in monitoring practice and P05 concedes the pull toward what is easier to evaluate |

**Step 2: worked judgment exercise with two defensible answers.**

Exercise. A generalist grantmaker has two candidates and one slot. Candidate one is a farm animal welfare monitoring program that would independently verify whether supply-chain commitments are being implemented, costing about two hundred thousand dollars per year. Candidate two is a careers organization that would place roughly fifty mid-career professionals into high-impact roles, costing about one million dollars per year. Which clears the bar, and on what does the answer depend?

- Answer A: fund the monitoring program. The persistence parameter in the farm animal welfare projection is the single most uncertain quantity in either portfolio, and a two-hundred-thousand-dollar measurement grant buys information that changes the valuation of a decade of existing commitments; that is a classic inframarginal move in the sense P06 describes, where the funder helps create something that then makes later grants far above the bar. This answer depends on believing the existing commitment portfolio is large enough that a persistence estimate has portfolio-wide value.
- Answer B: fund the careers organization. The placement program produces transitions now, and the information about persistence is only useful if someone acts on it, whereas the placement program's transitions are the welfare-relevant event itself. This answer depends on the persistence discount being small enough that the projected ten billion animals are not substantially discounted, in which case the farm animal welfare portfolio is already very strong and additional measurement has low marginal value compared with a new productive program.

The two answers differ on the value of information, not on values. A third answer also has to be considered and is worth naming: they are not mutually exclusive at these ticket sizes, and the corpus's own scale evidence makes that relevant. P02's Effective Careers RFP average grant is about $467,000, while P05's 2025 funding is nearly three times prior years, so the binding constraint at this ticket size is likely to be grantmaker time rather than dollars, which makes the decision about which intervention to investigate rather than which to buy.

**Step 3: why cross-area ranking is blocked at the currency step, and how to say so cleanly.**

```
import numpy as np

# Invented units. One unit of improved farm-animal life is assumed to be worth
# w units of one high-impact career transition. The exercise shows the *shape*
# of the comparison, not a value judgment.
faw_expected_animals = input_state.get(
    "faw_expected_animals", 10_000_000_000 * 0.5   # headline times a persistence parameter
)
careers_expected_transitions = input_state.get(
    "careers_expected_transitions", 1000.0         # one organization over one year
)

weights = np.logspace(-9, 0, 400)                  # plausible moral-weight range
faw_value = faw_expected_animals * weights
```

```
careers_value = np.full_like(weights, careers_expected_transitions)

# The weight at which the two are equal:
w_star = careers_expected_transitions / faw_expected_animals

output_state["weight_at_indifference"] = w_star
output_state["faw_expected_animals"] = faw_expected_animals
output_state["careers_expected_transitions"] =
    careers_expected_transitions
output_state["log10_w_star"] = float(np.log10(w_star))
```

The result is a threshold, not a ranking. Below the indifferent weight, the careers program values higher; above it, the animal welfare program does. Everything interesting about the comparison lives in the weight, and the corpus supplies no such weight: P03 supplies a human-income conversion in CG dollars, and nothing in the corpus converts animal welfare into the same unit. So the correct interview answer is that comparing these areas requires an explicit moral weight, and that a generalist who pretends otherwise is smuggling a value judgment into an empirical claim.

**Step 4: the generalist habits this brief is meant to install.**

- State the register before quoting a ratio: dollars moved, dollars of modeled social benefit, or welfare units are three different claims that the corpus renders with the same words.
- Convert conditionals into parameters: when a number is stated conditionally, name the condition and treat it as the uncertain quantity worth buying evidence about.
- Prefer designs with comparison groups over before-and-after counts, in every cause area, because counterfactuality and persistence are the shared load-bearing problems.
- Attach provenance: a P07 suggestion is a named individual's or team's judgment, an RFP reflection is a fund-level document, and a projected ten billion is a conditional claim. Do not flatten them into one voice.
- Do not treat a single case as a base rate: a famous early grant that worked, such as the small first grant to what became MATS, is one observation from a portfolio, and it is the same shape of evidence as the Manifund regrant that grew fifty-three fold. Both are tails, and a tail is evidence about the distribution only if you also know the denominator.

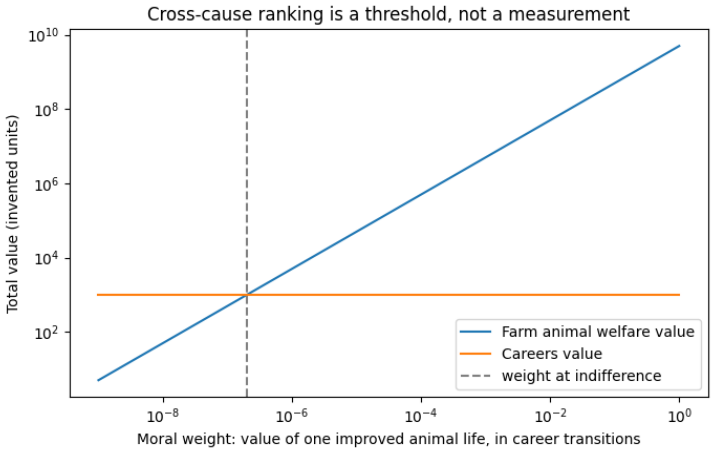


Figure 40: Graphical analysis for Step 1.

# Analyze abundance, growth and catch-up growth after AI.

## Question

Analyze abundance, growth and catch-up growth after AI as a Coefficient Giving cause area. Produce a briefing that explains what growth is doing inside CG, why the catch-up-growth argument (P03) is a different object from domestic abundance, which assumptions in that argument are load-bearing, what the main objections are, and how the recommendation changes when the adoption lag $D$ is 20 rather than 10.

Constraints: do not invent a grant, a portfolio or an institutional position. Separate source-supported claims from inferences and from proposed exercises. Treat the headline numbers as conditional on contested assumptions.

## Background

Source-supported facts. P06 (2025 Letter from the CEO) states that CG hired Matt Clancy to build the Abundance & Growth Fund with Dylan Matthews, Alex Armlovich, Willow Latham-Proença and Saloni Dattani, and that Justin Sandefur joined from the Center for Global Development to lead a new Global Growth Fund; the letter says CG is finally doing more to accelerate economic growth. P07 lists Center for Building in North America (CBNA) as an individual-donor suggestion in the abundance and growth area, and states explicitly that suggestions are individual staff views, not an institutional position.

P03 (Rafael Latham-Proença, "The Staggering Welfare Consequences of Catch-

Up Growth After AI") is a modeled welfare argument, not a fund announcement. It assumes AI accelerates frontier growth (the "detachment window") while poor countries adopt productive AI with a lag $D$, and concludes that the GDP gap widens during that window. Its central calibration is takeoff length $T = 15$ years, lag $D = 10$ years, full-automation growth $g_{\text{full}} \approx 18\%$, a 25-year window and a 4% discount rate. It reports average developing-world income about a fifth of the no-lag counterfactual, a present-value income shortfall of about $1.2 quadrillion, a welfare gap of about 6.1 quadrillion CG dollars, roughly 1.86 billion lives-equivalent, and per-year figures of about $42.6 trillion in present-value income or CG$381 trillion, both attached to roughly 115 million lives.

P03 also states its own reversal conditions and its critics' objections, and says the author's day job is making grants to accelerate catch-up growth. N14 (Centre for Effective Altruism) is the adjacent source on why technically sound policies fail without political feasibility, coalitions and implementation work.

## Sub-Question 1: What is growth as a CG cause area, and why are "abundance" and "catch-up growth" different objects?

P06 places two growth efforts inside CG: the Abundance & Growth Fund and a new Global Growth Fund led by Justin Sandefur. Both are called growth. An evaluator's first move is to refuse to treat them as one cause, because they differ on object, jurisdiction, lever, evidence and numeraire.

Abundance is domestic supply-side reform: housing and construction cost, where the wins are legislative (City of Yes, SB 79, hundreds of thousands of units) and the analysis is regulatory (CBNA's elevator and single-stair work). Catch-up growth is a conditional welfare claim about convergence after a frontier AI takeoff. Their only shared feature is the word "growth"; operationally they share almost nothing, including the timeline over which success would be observed.

*Note the citation discipline. P06 is the institutional source for both funds. P07's CBNA entry is an individual staff suggestion and says so. P03 is one grantmaker's model published as research, not a fund strategy or a commitment. Merging them into a single "CG growth voice" would be exactly the flattening the brief forbids.*

### Step 1

Separate the two strands along the dimensions that change a funding decision, not the ones that sound similar.

Table 77: *Two strands that share a word and little else. The numeraire row is the one that governs cross-cause comparison.*

| **Dimension** | **Abundance & Growth (domestic)** | **Global Growth / catch-up** |
|---|---|---|
| Object | Housing supply and construction cost; regulatory barriers | Convergence of non-frontier economies after a frontier AI takeoff |
| Jurisdiction | US states and cities | Low- and middle-income countries outside the frontier |
| Lever | Legislative reform, research, advocacy | Economic policy that changes the adoption lag D |
| Evidence cited | City of Yes, SB 79, 100k+ units; CBNA reports | P03 model: ~5x income gap, ~1.86B lives-equivalent |
| Numeraire | Dollars moved, units enabled, cost per unit | CG welfare dollars; lives-equivalent |
| Relation to the GCR mission | Essentially orthogonal | Explicitly ranked after preventing AI catastrophe, conditional on TAI |
| Institutional status | Existing fund (P06) | New fund led by Sandefur (P06); P03 is one author's model |

**The generalist's load-bearing observation: CG already operates at least three numeraires — dollars moved (P01), multiples of giving or social return (P02, P05), and CG welfare units with lives-equivalent (P03). A ranking across causes is only meaningful once you name the numeraire, because the same grant can rank first in one and third in another.**

## Sub-Question 2: Which assumptions in the catch-up argument are load-bearing, and is the reported arithmetic internally consistent?

P03's argument is a chain. Frontier takeoff raises the frontier growth rate; non-frontier economies adopt with a lag; during the detachment window the lag is a growing gap; the gap is welfare-valued; the welfare value is compared with the cost of other philanthropy. Each link has a parameter, and a grantmaker should

know which parameter moves the conclusion, which is merely descriptive, and which is a presentation choice.

$$g_F(t) = G + b(t), \qquad g_N(t) = G + b(t - D), \qquad b(t) \geq 0 \text{ rising in } f(t)$$

The mechanism is that growth rises with the automated share $f$, and $f$ follows an S-curve over $T$ years. The non-trivial consequence is that a fixed lag does not shrink the gap; it converts a gap in years into a growing gap in GDP, because the leader is on a steeper part of the curve.

**Step 1**

Rank the assumptions by how much of the conclusion they carry. The ranking is the deliverable; the list below is only the raw material.

Table 78: *Assumption ledger. Read the rightmost column first: it is the finding. Note that two of the author's own concessions (fixed lag too optimistic, lower $f_max/g_full$ for poor countries) both push the gap up, not down.*

| **Assumption** | **Source status** | **What it controls** | **If it fails** |
|---|---|---|---|
| Frontier AI takeoff is real | Contested: insiders 10–20%+; mainstream economists 0.1–1.5%/yr (P03 fn. 2) | Whether the AI-takeoff-specific divergence argument applies | This conditional argument loses its premise. Ordinary catch-up-growth benefits and any ranking against GCR require a separate assessment. |
| Follower is the same S-curve shifted by D | Author concedes probably too optimistic (Andersen comment) | Sign and shape of the gap | If poor countries get lower $f_max$ or $g_full$, the gap is larger, not smaller |
| Broad physical automation (high $f_max$) | Author concedes; "promises a future post about robots" | Magnitude of the welfare case | Case becomes "much less important" |

| Assumption | Source status | What it controls | If it fails |
|---|---|---|---|
| Takeoff length T = 15 (author's central guess) | Author's judgment between illustrative T = 5 and T = 30 scenarios; 15 is not their arithmetic midpoint. | Length of the detachment period | Scales the gap roughly with window length |
| $g_f ull \approx 18\%$/yr after full automation | Korinek–Suh 2024; author says not maximally aggressive | End-of-window growth and gap size | Lower $g_f ull$ shrinks the estimated gap |
| Capital accumulation underneath rising f | Author flags in fn. 8 and technical note | Whether rising f converts to growth at all | No growth acceleration; f rises without output gain |
| D is constant and does not grow | Directly disputed by Andersen; author concedes | Whether the lag is frictional or structural | Larger or growing D changes magnitude and the tractability question |
| Window 25 years, discount rate 4% | Presentation choices (calculator parameters) | All present-value and per-year figures | Headline scales; see the T + D = window point below |
| Post-window world is not "convergence" (Future 4) | Author enumerates four futures and says Future 4 is required for the blip objection | Whether the window matters for the recommendation | Gap is temporary and the window is a blip |
| A grant can change D at acceptable cost | Author explicitly does not know the policies | Whether the number becomes actionable | Important but not fundable yet — the N14 feasibility problem |

**The second concession is the subtle one. "Optimistic" and "conservative" are not the same axis when the model is conditional. A fixed lag is optimistic about whether the follower completes the catch-up,**

**but the gap it produces is a lower bound relative to a model where the follower never completes. That is why the argument's strength is not monotone in the plausibility of its assumptions.**

**Step 2**

Before quoting any headline, reconcile the numeraires. P03 attaches two different dollar figures to the same "115 million lives per year": about $42.6 trillion in present-value income and CG$381 trillion in welfare units. Those cannot be the same quantity unless an income-to-welfare conversion is doing work. The code below performs the consistency checks an evaluator would run by hand.

$$\text{CG welfare rate: } 1\% \text{ income gain} \times 1 \text{ person} \times 1 \text{ year} = \text{CG\$500}$$

Two checks are worth separating. First, whether the welfare totals are internally consistent with each other. Second, whether the income-to-welfare multiplier is the same at the per-year margin and at the window margin. If the first passes and the second fails, the totals are fine and it is the per-year figure that is stated on a different basis.

## Sub-Question 3: How much of the headline survives a change in D, and what happens to the recommendation when D is 20 rather than 10?

P03's central scenario is $T = 15$, $D = 10$ and a 25-year observation window. In the stylized logistic example below: $T + D = 15 + 10 = 25$. *At that date the toy follower curve reaches its 95% automation benchmark. Logistic automation approaches 100% asymptotically, so this equality does not imply full automation or an abrupt change in the model.*

Increasing D beyond 10 moves the toy follower's 95% automation benchmark beyond the observation window. This can change measured outcomes substantially, but continuously; it does not create a mathematical knife edge or establish a corresponding threshold in P03's original model.

**Step 1**

Build the smallest model that reproduces the qualitative claims: a logistic automation curve for the frontier, and the same curve shifted right by D for the follower. Growth is increasing in f, so the log-income gap over the window is the area between the two curves, scaled by a growth-elasticity $\beta$. Calibrate $\beta$ so that D = 10 reproduces the reported ~5x income gap, then read off other lags.

$$f(t) = \frac{1}{1 + \exp[-k(t - t_0)]}, \qquad k = \frac{2 \ln 19}{T}, \qquad t_0 = \frac{T}{2}$$

$$\Delta(D, W) = \beta \left[ \int_0^W \sigma(k(t - t_0))\, dt - \int_0^W \sigma(k(t - t_0 - D))\, dt \right],$$

$$\frac{\text{income with no lag}}{\text{income with lag}} = e^{\Delta}$$

Two properties are built into this specification and must be stated: the steepness k is fixed by T, and $\beta$ *is a stylized growth elasticity. The source does not give the S-curve shape, so this is a sensitivity exercise, not a re-derivation of P03.*

### Step 2

Show the two paths and the gap they enclose. The vertical line at $T + D = 25$ marks the end of the observation window. In this toy model, the D = 10 follower is 95% automated then, while the D = 20 follower is about 27% automated. Neither curve reaches 100% at a finite date.

*Read the figure against the reported result rather than in place of it. The shape is the finding; the exact ratio is a function of the stylized steepness.*

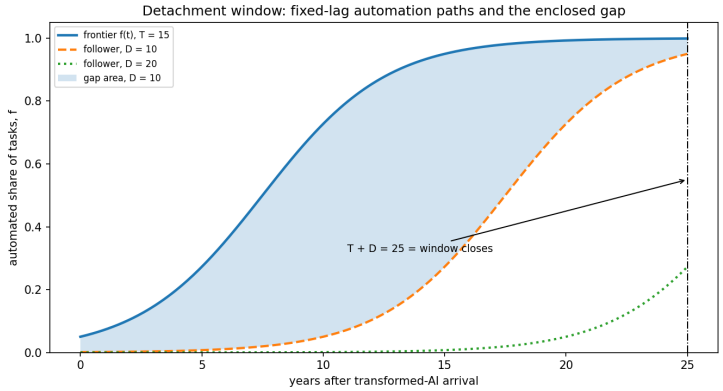


Figure 41: Graphical analysis for Step 2.

### Step 3

What the exercise shows, reported as an evaluator would report it. In this toy calibration, the log-income gap is convex for $0 \leq D < 5$ and concave for $D > 5$. Doubling D from 10 to 20 raises the log gap by roughly 1.7x; the corresponding income ratio rises from 5x to about 15.4x, a roughly 3.1x increase. At sufficiently large D, the follower makes little progress inside the observation window and further increases in D have diminishing effects.

Table 79: *Stylized sensitivity from the model above. The classification column is the finding: the case strengthens in magnitude as D rises, and the last row is the only quantity a grantmaker can purchase.*

| **Lag D** | **Income ratio vs no-lag** | **Reading** |
|---|---|---|
| 5 | ~2.3x | Follower on the steep part with the frontier; small gap |
| 10 | ~5x (calibrated to P03) | Follower reaches the toy model's 95% benchmark at t = 25; this is not full automation or a discontinuity. |
| 20 | ~15x | Follower's takeoff largely truncated by the window |
| Very large / structural | ~18x (ceiling) | Follower effectively never completes; gap saturates |
| Moving D down by 1 year, at D = 10 | ~+17.1% end-window income (D = 10 to D = 9) | This marginal quantity, not the total gap, is what a grant could buy |

**The recommendation change is paradoxical and worth saying out loud in an interview. Raising D to 20 makes the welfare gap larger, so the argument for treating catch-up as important strengthens. But it also makes the lag more plausibly structural — institutions, energy, conflict, capital access — rather than frictional, so the amount of D that a grant can actually move probably shrinks. The value of an intervention is (value per year of D moved) x (years of D the intervention moves), not the whole gap. A 3x larger gap with a 3x less tractable lever is roughly a wash, and it moves the funding question from "is this big?" to "is there a lever?"**

*P03 licenses this reading directly: the author says his day job is making grants to accelerate catch-up growth, that policy can change D, and that he does not know which policies. A dossier that quotes the trillion-dollar figure without the lever caveat would be doing advocacy, not evaluation.*

## Sub-Question 4: What are the main objections, how would a grantmaker adjudicate them, and why does this matter to a GCR or generalist evaluator?

P03 organizes objections into AI-skeptic and AI-believer camps and answers each. The evaluator's job is not to pick a camp but to identify which objections are cruxes the funder could buy evidence about, which are worldview disagreements, and which (the most important one) the source itself leaves open: whether a grant can move D at all.

### Step 1

Lay the objections out with their responses and, separately, with what would move the evaluator. The third column is the one that turns a reading into a judgment.

Table 80: Objection ledger. The bottom two rows are the ones P03 does not own, and they are the ones a grantmaker is actually being interviewed about.

| **Objection** | **P03 response** | **What would move an evaluator** |
|---|---|---|
| Skeptic: the takeoff never comes (mainstream estimates 0.1–1.5%/yr) | Believers' recent track record beats skeptics'; the post is conditional on insiders being roughly right | A world model of AI growth effects that survives both camps; not a blog-post contrast |
| Skeptic: the assumed takeoff is too high (FRI ~5% by 2050; IMF SSA ~4% over 10 yrs) | Those scenarios are not full automation; slide $f_max$ down to reach them; 18% may even be too low | The parameter slider agreement is a concession, not a rebuttal — most real forecasts sit in the low-$f_max$ region |
| Skeptic: only cognitive tasks automate, so $f_max$ is low | Author concedes: without broad physical automation the main point "becomes much less important" | Cheap evidence exists: robot adoption and cost curves by sector and country |
| Skeptic-believer: D is too low and may grow; historically follower curves did not complete | Author concedes the shifted-S-curve is too optimistic; lower $f_max$ or $g_full$ for poor countries would strengthen the point | Most decision-relevant crux, but it is a structural claim about institutions, not a frictional-lag claim |

| Objection | P03 response | What would move an evaluator |
|---|---|---|
| Believer: AI itself shrinks D by lowering technical barriers | Hopeful but not enough to overcome the four barriers | Evidence from firm-level AI adoption in low-income production |
| Believer: trade drives catch-up | Real but narrow; physical automation makes frontier manufacturing more, not less, competitive; exports are extractives and tourism | Sectoral export data under automation scenarios |
| Believer: frontier regulation raises frontier costs, compressing D | Race dynamics between US, China and Europe dominate | Evidence on enforcement capacity, not statute text |
| "It's just a blip": abundance arrives, income differences stop mattering | Only holds in Future 4 (convergence); access needs income; lock-in makes early gaps consequential | A defensible disagreement about the post-window world, not resolvable with the current corpus |
| Evaluator's own: can a grant change D? | Author: policy can change D, but "I won't pretend I know exactly what policies" | A concrete lever with a measured effect on adoption intensity; until then the number is conditional |
| Evaluator's own: attribution and counterfactuality | Not addressed; P03 says it will learn and publishes a calculator | The same counterfactuality adjustment P02 and P05 apply to ROI, applied to D |

**Adjudication rule: separate the empirical part from the structural part of each objection. "Will physical tasks automate?" is empirical and buyable. "Will the follower complete the curve?" is largely structural and is where the argument's tractability lives. "Should we discount the far future?" is a values disagreement and is not a crux.**

**Step 2**

Connect the argument to the adjacent source. P03's "What to do about it" section is exactly the region N14 maps: policies on issues with no academic consensus are premature, and the missing work is response-identification, coalition-building and political-feasibility analysis before advocacy. "I don't know the

policies" is not a gap in P03's diligence; it is the state of the field that N14 predicted.

- Relevance to the GCR mission: P03 itself says ensuring developing-country growth stays close to the frontier might be the second most important task in human welfare, after preventing AI-induced catastrophe — and flags "all huge ifs" in a footnote. The evaluator should hold it as a conditional second priority under a stated worldview, not as a co-equal GCR thesis.
- Institutional difference to preserve: P06 describes a diversified AI strategy with no single theory of change, and P03 is one grantmaker's model. Abundance is a different fund with a different lever. Collapsing these into "CG believes growth is the fix" would misstate the evidence.
- Tool transfer: the evaluator's instruments are the same across causes — theory of change, counterfactuality, cost-effectiveness, scale, tractability. What changes is the numeraire. P03 supplies a welfare numeraire; P02 and P05 supply a multiple-of-giving numeraire; P01 supplies amounts committed. Ranking catch-up growth against bed nets requires picking one and defending the conversion.
- Interview framing, calibrated and short: "P03 is a conditional welfare argument, not a grant. Its magnitude is large and its own concessions push it larger, but its load-bearing assumptions are the frontier takeoff, the fixed shifted lag, and broad physical automation — and the one that decides whether it is fundable is whether any grant can move D. I would fund the response-identification work that N14 says is missing before funding implementation."

*Open uncertainties to carry forward: the exact income-to-welfare multiplier is not stable across margins (the code above); the discount rate, window and takeoff length are presentation parameters that scale the headline; D, $f_max$ and $g_full$ are contested; and whether the post-window world is convergence, plateau, fade or singularity determines whether the window matters at all. None of these are resolved by the corpus, and a briefing that presents them as settled would be the failure mode.*

# Reason about cross-cause prioritisation and worldview diversification.

## Question

Produce a memo that compares marginal returns across Coefficient Giving cause areas and explains marginalist funding, worldview diversification, general-manager ownership, and the tensions among them. Use source-supported claims from the supplied corpus, mark inferences, and include a worked 10 million dollar allocation exercise with at least two defensible answers and explicit revision conditions.

## Background

Coefficient Giving is a multi-fund grantmaker. In 2025 it directed over 1 billion dollars, increased giving by more than 50 percent, and now expects higher budgets across funds (P06). Its major cause areas in the supplied corpus include AI safety and security; biosecurity and pandemic preparedness; global health and wellbeing; farm animal welfare; effective giving and careers; abundance and growth; and science and global health R&D (P01, P04, P05, P06, P07, S01, S04).

The evaluator role may test generalist judgment across these areas, not only catastrophic-risk reasoning. The relevant question is not whether one cause area is good in the abstract but whether the next dollar has a higher marginal return in one area than another, and what value assumptions make that comparison possible (P06).

P06 describes a historical marginalist mode: each additional dollar should produce roughly the same amount of good if allocated to any fund, because if it does not, resources should shift toward the higher-return area until they do. P06 also records zero-sum budget decisions across worldviews and funds, a continued shortfall of funding relative to the opportunity set, and a shift toward more ambitious goals and general-manager ownership as budgets grow.

P07 says individual donor suggestions are not institutional views and are not necessarily the best options across areas. P08 says CG focuses on the highest-stakes risks while supporting a wide range of experts, including those who disagree with CG and each other. D01 requires calibrated confidence and transparency about the kind of support behind major claims. P03's coda models how assumptions and values change prioritisation and warns against both AI-skeptic and AI-believer errors.

This memo separates three layers. Source-supported claims are attributed to P01, P02, P03, P04, P05, P06, P07, P08, S01, S04, and D01. Inferences are marked as inferences. Proposed exercises are marked as proposed and are not evidence about Coefficient Giving's actual allocation.

## Sub-Question 1: Marginalist Funding and the Cross-Cause Comparison Problem

What operational rule does marginalist funding imply, and what must be true before marginal returns can be compared across cause areas such as AI safety, biosecurity, global health, farm animal welfare, effective giving and growth?

Answer in the form of a memo section: state the rule, build a comparison scaffold, and identify the conversion problem that makes the comparison non-trivial.

### Step 1

Source-supported. P06 describes a marginalist mode of grantmaking: think hard about how to equalise marginal returns within and sometimes across cause areas. The idea is that each additional dollar should produce roughly the same amount of good if allocated to any fund. If it does not, shift resources toward the higher-return area until it does. P06 also records zero-sum budget decisions across worldviews and funds, and says funding still falls short of the opportunity set.

$$r_i = \frac{\partial W}{\partial x_i}$$

At an interior optimum with a fixed budget B, the first-order condition is $r_i$ = lambda for every funded area i, and $r_i$ <= lambda for areas that receive zero. This is the equal-marginal-return condition. It is a formal rule, not a substantive finding.

Inference. The rule does not tell an evaluator what W is, how to measure $r_i$, or whether the opportunity set is smooth. P06 notes that many of its best grants have been deeply inframarginal: CG had to recruit a founder or effectively incubate an organisation, and once something fundable existed the eventual grant was far above the bar. Those are lumpy, not marginal, decisions. Marginalist funding is therefore a default discipline, not a complete decision rule.

### Step 2

To compare across cause areas, build a table that keeps four things separate: the primary outcome metric, the evidence anchor, the counterfactuality problem, and the marginal return driver. Comparing headline ROI numbers across areas without this scaffold is the most common cross-cause error.

Table 81: Source anchors are illustrative. P07 is a set of individual staff suggestions, not an institutional ranking.

| Cause area | Primary outcome metric | Evidence anchor | Counterfactuality problem | Marginal return driver |
|---|---|---|---|---|
| AI safety and security | Catastrophic risk reduction | P01, P06, P08, D04 | Whether an additional grant changes frontier-lab behaviour, governance trajectories, or talent supply | Absorptive capacity, talent, compute, and research paradigms |
| Biosecurity and pandemic preparedness | Pandemic risk reduction | P01, P04, P06 | Whether preparedness spending reduces probability or only consequences | Detection, bioharden-ing, PPE, rapid medical countermeasures |
| Global health and wellbeing | Lives saved and income gains | P06, S01, S04, P07 | Whether CG funding displaces other donors or adds new spending | Cost per life saved, evidence quality, room for more funding |
| Farm animal welfare | Animals improved | P06, P07 | Moral weight of animal suffering and durability of corporate commit-ments | Corporate campaigns, supply-chain leverage, neglected species |

| Cause area | Primary outcome metric | Evidence anchor | Counterfactuality problem | Marginal return driver |
|---|---|---|---|---|
| Effective giving and careers | Dollars moved and career transitions | P02, P05, P07 | Whether movement is additional and whether counterfactuality adjustments hold | ROI of effective giving, career-transition rates, incubation pipeline |
| Abundance and growth | GDP growth and housing supply | P03, P05, P06, P07 | Long causal chain from policy to growth and attribution across many actors | Frontier growth, catch-up growth, regulatory reform |
| Science and global health R&D | Vaccines and repurposed drugs | S04, P07 | Market failure and clinical trial risk; whether philanthropy is the binding funder | Neglected disease burden, tractability, cost per trial |

The table exposes a non-trivial point. The outcome metrics are not commensurable. Lives saved, animals improved, dollars moved, GDP growth, and catastrophic risk reduction cannot be added without weights. The table is therefore a comparison scaffold, not a decision.

**Step 3**

Source-supported. P03 supplies one CG-adjacent conversion: a 1 percent income increase for one person for one year is valued at CG$500, and a present-value income shortfall of about 42.6 trillion dollars is treated as equivalent to saving roughly 115 million lives under the CG methodology. P02 reports a weighted mean modelled social return on investment of roughly 5,500x for the effective careers RFP, but calls the estimate early and uncertain. P05 reports an effective giving portfolio ROI of roughly 5 to 6x. These numbers are not the same object. One is a social return on donated dollars, one is an income-equivalent welfare conversion, one is a portfolio ROI, and one is a lives-saved equivalence.

$$W = \sum_i w_i \cdot O_i, \quad r_i = w_i \frac{\partial O_i}{\partial x_i}$$

Inference. Equalising $r_i$ requires $w_i$, the welfare or moral weight attached to each outcome. Those weights are exactly what worldview disagreement disputes. A table of ROIs cannot settle cross-cause prioritisation without an explicit or implicit set of $w_i$. D01 requires the evaluator to state how confident they are in those weights and what kind of support they have for them.

What would change the allocation. A revision in any of $w_i$, $O_i$, the shape of diminishing returns, or the budget B changes the cross-cause allocation. The memo should therefore state the weights and the marginal-return assumptions rather than presenting one allocation as obviously correct.

## Sub-Question 2: Worldview Diversification versus Focused Cause Prioritisation

What is worldview diversification, how does it differ from focused cause prioritisation, and when should an evaluator recommend each?

Explain the value-uncertainty logic, the empirical-uncertainty logic, and the source-supported tension with marginalist funding.

### Step 1

Source-supported. P06 records zero-sum budget decisions across worldviews and funds. Worldview diversification is not the same as empirical hedging. Empirical hedging spreads bets because you are uncertain about facts within a shared value framework. Worldview diversification spreads allocations because you are uncertain about the value framework itself.

$$E[U(a)] = \sum_w p_w u_w(a)$$

Inference. If the probabilities $p_w$ over worldviews are known and utility is linear in the allocation, the optimal policy is to focus on the allocation that maximises expected utility. Diversification is optimal under ambiguity aversion, correlated errors, or imprecise $p_w$. It buys robustness and option value. It is a second-best response to unresolved normative uncertainty, not a denial that some worldviews are more defensible than others.

Source-supported. P08 says CG aims to support a wide range of experts, including those who disagree with CG and each other. P06 says CG pursues a diversified approach to AI giving with no single theory of change, and warns against the strategic-funder trap of treating grantees like subcontractors. Those are institutional expressions of openness that resemble worldview diversification within a cause area.

**Step 2**

Compare the two postures directly.

Table 82: Examples are source anchors, not institutional allocations.

| Dimension | Focused cause prioritisation | Worldview diversification |
|---|---|---|
| Objective | Maximise expected value under the best current value framework | Hedge against unresolved value uncertainty and correlated worldview error |
| When preferred | Value framework is stable and one area has sharply higher marginal return | Value uncertainty is high and allocations are correlated with one contested framework |
| Main risk | Overconfidence in contested moral weights and cause-level estimates | Diluting resources across views one has little reason to take seriously |
| Evidence requirement | Cause-level cost-effectiveness and counterfactuality estimates | Explicit weights across worldviews plus confidence in each |
| CG example | P06 marginalist equalisation across funds | P08 support for experts who disagree; P06 diversified AI approach |

Source-supported. P06 says the trade-offs that drive a focus on equalising marginal returns remain, funding still falls short of the opportunity set, and there are still zero-sum budget decisions across worldviews and funds. P07 says individual donor suggestions are not institutional views and are not necessarily the best options across areas. Those statements cut against treating any single list or fund as the whole of CG's view.

**Step 3**

- Diversify when value uncertainty is high and errors across worldviews are correlated with one's own institutional position.
- Focus when a worldview dominates after reflective equilibrium and one area has sharply higher marginal returns at the relevant budget.

- Do not diversify across worldviews you have no reason to take seriously; that is moral indifference, not uncertainty hedging.
- Use marginalist logic within each worldview and then apply a meta-weight across worldviews; keep the meta-weights visible.
- Avoid the strategic-funder trap: diversification should increase openness to grantees' local context, not impose CG's worldview on them (P06).

Non-trivial jump. Worldview diversification and marginalist funding can conflict because marginalist funding assumes a single W, while worldview diversification assumes multiple W. The resolution is not to abandon either. It is to make the meta-welfare function explicit, state confidence in it, and then apply the equal-marginal-return logic under that meta-function. D01's reasoning-transparency norms apply directly.

## Sub-Question 3: General-Manager Ownership and the Tension with Marginalist Funding

What is general-manager ownership in CG's own framing, and how does it create a tension with marginalist funding?

State the conditions under which proactive ownership is justified and the conditions under which it is not.

### Step 1

Source-supported. P06 describes a shift as budgets grow: away from optimising marginal trade-offs and toward more ambitious goals. It praises the general-manager model: take full responsibility for a problem, recruit a founder, or effectively incubate an organisation. P06 cites Nan Ransohoff's argument that more of the world's important problems should have general managers, and gives Andrew Snyder-Beattie's biosecurity work and Lewis Bollard's farm animal welfare work as examples.

Source-supported. P01 says CG is capacity-constrained and staff time is the main thing preventing scaling. P01 also describes Project Tailwind, a call for founders to start ambitious AI safety initiatives. P06 says helping create new grantees is far more labour-intensive per dollar moved at the beginning, but can pay bigger dividends in terms of subsequent capacity and impact later. P06 also warns against the strategic-funder trap of thinking CG has all the answers and slotting grantees in like subcontractors.

### Step 2

The tension is structural, not personal.

Table 83: The table contrasts ideal types. P06 says the right balance varies across focus areas and individuals.

| Dimension | Marginalist funder | General-manager owner |
|---|---|---|
| Posture | Respond to opportunities and select the highest marginal return | Create opportunities and own the problem end-to-end |
| Decision rule | Fund if the grant clears the cost-effectiveness bar | Recruit founder, incubate organisation, build capacity, then fund at scale |
| Time horizon | Near-term marginal dollar | Longer-term inframarginal win |
| Use of grantmaker time | Lower per dollar moved for existing grantees | High per dollar moved at the beginning |
| Main risk | Missing inframarginal wins that require creation | Imposing CG's vision, treating grantees as subcontractors, and consuming scarce staff time |
| Source anchor | P06 marginalist mode | P06 general-manager discussion; P01 Project Tailwind |

Source-supported. P06 wants both: the ambition and ownership of general managers without losing the curiosity and humility of marginal funders. That is the central tension for an evaluator. P06 says the right balance varies across focus areas and individuals, so an interview answer that resolves the tension with one universal rule is likely to be less calibrated than one that states the conditions.

**Step 3**

Inference. A decision rule for when general-manager ownership is justified.

- The problem is unowned or the opportunity set lacks organisations that can absorb capital.
- Founder supply is the binding constraint and CG has comparative advantage in recruiting or incubating.
- The expected inframarginal gain is large relative to the opportunity cost of grantmaker time.
- Failure is tolerable and can be staged through pre-seed, seed and scale tiers such as those in P01's Project Tailwind.

- CG's involvement adds capacity rather than substituting for stronger grantee local knowledge.

Counterconditions. Do not pursue general-manager ownership when existing organisations can absorb marginal dollars at a higher return, when the grant-maker's time is the scarce input for a higher-return opportunity, when CG would impose its own vision, or when the causal chain is too long to learn from. P06's warning about the strategic-funder trap is the discipline that keeps ownership from becoming overreach.

## Sub-Question 4: Worked Exercise: Allocate a Marginal 10 Million Dollars Across Three Cause Areas

Allocate an additional 10 million dollars across AI safety and security, biosecurity and pandemic preparedness, and global health and wellbeing. Give at least two defensible allocations, show the marginal-return comparison, and state what would change the allocation.

Treat the numbers as a proposed exercise, not as Coefficient Giving institutional positions. Use a common welfare unit and explicit assumptions.

### Step 1

Set up a simple marginal-return model. Let $x_i$ be millions allocated to cause i. Assume diminishing marginal returns of the form $r_i(x_i) = a_i$ exp(-$x_i$ / $b_i$), where $a_i$ is the marginal return at zero and $b_i$ controls how quickly returns fall. The budget is 10 million dollars.

$$r_i(x_i) = a_i e^{-x_i/b_i}, \quad \sum_i x_i = 10$$

At an interior optimum, equal marginal returns give $x_i = b_i$ (ln $a_i$ - ln lambda), with ln lambda = ($sum_i$ $b_i$ ln $a_i$ - B) / $sum_i$ $b_i$. The parameters are illustrative. They are chosen to make the exercise concrete, not to claim that these are CG's estimates.

Base parameters. AI safety and security: a = 120, b = 6. Biosecurity and pandemic preparedness: a = 80, b = 5. Global health and wellbeing: a = 60, b = 8.

### Step 2

Compute the equal-marginal-return allocation under those parameters.

Table 84: Illustrative equal-marginal-return allocation. Units of marginal return are arbitrary welfare units per million dollars.

| Cause area | $a_i$ | $b_i$ | Allocation in millions | Marginal return at optimum |
|---|---|---|---|---|
| AI safety and security | 120 | 6 | 5.55 | 47.6 |
| Biosecurity and pandemic preparedness | 80 | 5 | 2.60 | 47.6 |
| Global health and wellbeing | 60 | 8 | 1.85 | 47.6 |

This is defensible answer A. It follows the marginalist rule literally under the stated parameters. It does not mean global health is unimportant. It means that under these assumptions global health has a lower marginal return at zero and a larger diminishing-return scale, so it receives the residual after the two higher-a areas. The allocation is sensitive to $a_i$ and $b_i$. If $a_G H$ rises to 100 while other parameters hold, global health receives a materially larger share.

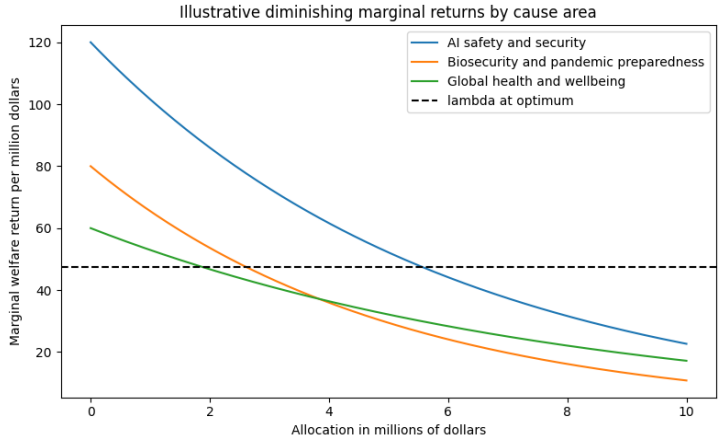


Figure 42: Graphical analysis for Step 2.

**Step 3**

Defensible answer B: a worldview-diversified allocation. If the evaluator treats the moral weights $w_i$ as unresolved, or treats the AI safety estimate as more

speculative than the global health evidence, diversification is warranted. A reasonable diversified allocation is 3.5 million to AI safety, 3.0 million to biosecurity, and 3.5 million to global health. A second variant is 3.0 million to AI safety, 3.0 million to biosecurity, and 4.0 million to global health.

Table 85: All allocations are proposed exercises, not CG institutional positions.

| Scenario | AI safety | Biosecurity | Global health | Key assumption | What would change it |
|---|---|---|---|---|---|
| Marginalist base | 5.55 | 2.60 | 1.85 | Parameters above; single welfare function | Updates to $a_i$ or $b_i$ |
| GCR-focused | 5.0 | 3.0 | 2.0 | Highest-stakes catastrophic risk reduction | Evidence that biosecurity or global health marginal returns are higher |
| Worldview-diversified | 3.5 | 3.0 | 3.5 | Unresolved moral weights and correlated worldview error | Reflective equilibrium or new moral argument |
| Evidence-weighted | 3.0 | 2.0 | 5.0 | Global health has the strongest RCT evidence and counterfactuality | Improved counterfactuality estimates for AI safety or biosecurity |

Both answers are defensible because they rest on different but explicit assumptions. Answer A is stronger if the evaluator accepts a single welfare function and the illustrative parameters. Answer B is stronger if the evaluator takes nor-

mative uncertainty seriously or discounts the AI safety marginal return more heavily. D01 requires the evaluator to state which assumptions are doing the work.

**Step 4**

What would change the allocation.

- AI safety and security: a replicated uplift result or a control-protocol failure raises $a_i$; failed replications or sustained refusal rates lower it.
- Biosecurity and pandemic preparedness: evidence that screening, detection, or rapid medical countermeasures are newly tractable raises $a_i$; evidence that preparedness spending changes consequences more than probability lowers it.
- Global health and wellbeing: displacement evidence lowers the counterfactuality adjustment; a lower cost per life saved raises $a_i$.
- Worldview weights: reflective equilibrium or a new moral argument changes $w_i$ and therefore the meta-allocation.
- Budget: if the marginal budget rises above 10 million, equal marginal returns push allocations toward the lower-a areas; if it falls, the higher-a areas dominate.
- Grantmaker time: if the binding constraint is staff time rather than dollars, the allocation should shift toward instruments that lower application and evaluation cost.

Transparency condition. P07 warns that individual suggestions are not institutional views. P08 says CG supports a wide range of views. P06 says the right balance between marginalist funding and general-manager ownership varies. A calibrated memo should therefore present the allocation as conditional on stated assumptions, attribute source claims, and mark the rest as inference or proposed exercise.

# Understand policy process, political feasibility and implementation.

## Question

Build a policy-process brief that explains feasibility, coalitions, policy windows, implementation, and the specific reasons a technically sound proposal can die. The brief must be usable as interview material for an evaluator or grantmaker at Coefficient Giving, which means it has to connect a political-science account of the policy process to a funding decision about a specific proposal.

The step is deliberately narrow. N14 supplies the record of a real attempt to move policy in the UK: more than a dozen meetings with policymakers, hundreds of generated options, the overwhelming majority set aside, and a small number investigated in depth. N06 supplies the map of how a policy recommendation actually travels from an idea to a change in what an AI lab does. D04 supplies the criteria a funder will apply to a policy proposal. The task is to make these three sources speak to each other.

## Background

The role brief describes grantmakers who investigate funding opportunities, develop neglected projects, research programme strategy and historical cost-effectiveness, and compare opposing arguments. Policy is one of the portfolios where those skills are tested most sharply, because a policy proposal can be excellent on its own terms and still be worthless, and the evaluator has to be able to say why without retreating into either cynicism or advocacy.

D04 lists policy development as an eligible subject area and names a specific output standard: proposals developed in enough detail that they can be debated and implemented by policymakers, typically needing more development and debate than they currently have. That standard is the bridge to N14, which is a field report on exactly the work of taking a high-level idea and trying to get it debated and implemented, and on the reasons that work mostly failed.

*Two cautions govern everything below. First, N14 is a small-n organisational memoir, not a study: the sample is one organisation in one country over roughly a year, the rejected proposals are mostly not published, and the author says the lessons will be obvious to anyone with policy experience. Second, N06 is explicitly partial: it is built mainly on EU, US, OECD, France and G7 experience and on interactions with developers in the US, the UK and multinationals, and it leaves five macro-areas off the map. Every generalisation in this brief is therefore an inference from a small, situated, self-described partial record, and should be labelled as such in an interview.*

## Sub-Question 1: The Anatomy of Political Infeasibility: Four Rejection Mechanisms, Four Cost Bearers

N14 reports that the most common reason a proposal was rejected was that it was politically infeasible. Explain why that verdict is a composite rather than a single finding, identify which of its components a philanthropic funder can actually move, and say what follows for an evaluator reading a policy proposal.

### Step 1

Start with the word itself. To call a proposal **politically infeasible** sounds like a description of the proposal. It is not. It is a description of the relationship between the proposal and a particular set of decision-makers who bear particular costs. N14's own examples separate cleanly into mechanisms that have almost nothing in common except that the same phrase covers them.

The organising question is not how good the proposal is. It is: who bears the cost, and can that bearer be paid? **The first non-trivial move is to stop treating feasibility as a property of the proposal and start treating it as a property of the proposal-bearer pair.**

Table 86: Four mechanisms hidden inside the single verdict politically infeasible, with the cost bearer and the funder lever that acts on it. All N14 instances are verbatim from the rejected-policy list.

| Rejection mechanism | Who bears the cost | N14 instance | Remedy class | Funder lever |
|---|---|---|---|---|
| **Financial cost to government** | The treasury or a ministry budget line | Funding research to improve forecasting; increasing funding for research into global catastrophic risks | External or co-funding; revenue-neutral or self-financing design | Pay the cost so it never lands on the treasury. This is the most directly payable term. |

| Rejection mechanism | Who bears the cost | N14 instance | Remedy class | Funder lever |
|---|---|---|---|---|
| **Cost concentrated on an identifiable voter group** | A named bloc: producers of, or consumers of, the taxed or subsidised good | Taxing all non-free-range meat products; stopping subsidies of certain agricultural products | Compensation, phased introduction, or bundling with a visible offsetting benefit | Finance transition payments or a visible offset. Least amenable to a distant funder because the compensation has to be credible to the bloc. |
| **Political capital, weirdness, unpopularity** | The individual politician's reputation, attention and staff time | Disaster shelters to protect against global catastrophic risks, judged too weird; increased immigration, judged too unpopular | Reframing, bundling into a popular vehicle, reducing the sponsor's exposure | Supply the research, drafting, polling and legal review the politician would otherwise spend scarce staff time on. This lowers the sponsor's private cost without changing the policy's content. |
| **Blockability by organised interests** | The politician's floor coalition and legislative time | Setting minimum sizes for gestation crates for pigs, which could be blocked by special interests | Venue shift, counter-coalition, splitting or buying off the opposition | Fund counter-coalition infrastructure or an alternative venue where the opposition has fewer veto points. |

The inference a grantmaker should draw is uncomfortable. If the most tractable term is the treasury cost, then the funder's comparative advantage is strongest precisely where the proposal looks least strategic: paying for something a government should arguably pay for itself. And the term that most excites policy advocates, the strength of the argument, is not on the list at all, because N14 never reports a proposal failing on the quality of its reasoning.

This is the point an evaluator should carry into an interview. **A grant that improves the argument for a policy and a grant that lowers the cost of adopting it are different instruments aimed at different terms in the same decision, and the second is usually the one with traction when the first has already been tried.**

**Step 2**

The four mechanisms can be written as one decision condition. Let the politician adopt when

$$U = B_{\text{perceived}} - C_{\text{treasury}} - C_{\text{voters}} - C_{\text{capital}} - P_{\text{blocked}}\, L_{\text{defeat}} - U_{\text{alternative}} \;\geq\; 0$$

where $B_{\text{perceived}}$ is the believed benefit, the three $C$ terms are the costs the four mechanisms impose on the politician, $P_{\text{blocked}}\, L_{\text{defeat}}$ is the probability-weighted cost of a visible defeat, and $U_{\text{alternative}}$ is what the same political capital would have bought elsewhere. N14's own gloss on the third mechanism is exactly this last term: policymakers were unwilling to spend political capital because they had more promising uses of their time and money.

Now the sharpest thing N14 says, which is easy to read past. Both of the detailed failures the post investigates, an IPCC-like body for biotechnological risk and a parliamentary committee inquiry, were instruments for **communicating expert opinion**. Both are aimed at raising $B_{\text{perceived}}$. And N14 reports that policymakers were *more willing to accept that there was a problem than expected*: believing a problem exists was not sufficient to make anyone develop responses, and it was expected that the advocates would do that work.

So the failures do not show that consensus-communication tools are useless. They show that those tools were pointed at a term that was not binding, in a setting where the binding term was the **supply of ready responses**. That is the deepest move available in this material, and it requires holding two facts together: the instrument was reasonable in the abstract, and it was misallocated given the actual constraint.

*One can then rank the funder's leverage by term. This ranking is an inference from N14, not a claim N14 makes: treasury cost is directly payable; political-capital cost is partly purchasable by absorbing the staff work; blockedness is movable through counter-coalition funding; voter-group cost is the hardest to*

*buy from outside because the compensation has to be credible to the bloc being compensated.*

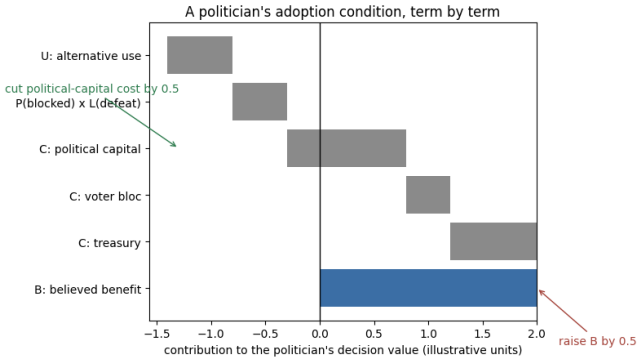


Figure 43: Graphical analysis for Step 2.

## Sub-Question 2: The Staged Model as a Multiplicative Chain: Locating the Binding Gate

N14 presents an adapted rational model of policy development as a sequence of stages and then uses it to argue that work at later stages is premature while earlier stages are unfinished. Read the model as a stage-gate product rather than a sequence, show why a single near-zero gate caps the value of the whole policy, and apply the diagnostic to the two documented failures.

### Step 1

N14's operative sentences are sequential in form: the community is at the stage of identifying the problem, so it would not be useful to shift attention to later stages; previously rejected policies can be picked up when a policy window opens; the intergovernmental-panel idea was an attempt to move from a coalition around the problem to a coalition around responses, and it failed because the earlier coalition was not there. Read as advice about ordering, this is unremarkable. Read as a constraint on value, it is a much stronger claim, and it is the one worth carrying into an evaluation.

The non-trivial jump: **the stages behave like factors in a product, not like steps in a queue.** A policy produces its effect only if every gate is passed, so the probability of effect is

$$V = \prod_{i=1}^{n} p_i, \qquad \frac{\partial V}{\partial p_i} = \prod_{j \neq i} p_j = \frac{V}{p_i}$$

The derivative carries the whole argument. **The value of one unit of improvement at gate i equals the product of every other gate**. So a proposal that is near-perfect at response identification and drafting but near-zero at political feasibility is worth almost nothing, and improving the parts that are already strong adds almost nothing. This is why N14's failures are not primarily failures of analysis.

There is a second, sharper consequence. Perfecting the weakest gate raises the product only to the product of the remaining gates. If a second gate is also very small, no amount of work on the first gate gets the policy above the second. The practical test is therefore not which gate is smallest, but whether there is a second gate within an order of magnitude of it. If there is, the correct action may be to stop repairing this policy and go find one whose gate profile has no pair of small factors, or to change the venue so that a different gate vector applies, which is the subject of the fourth subquestion.

Table 87: Both failures are gate failures rather than merit failures. The first gate is substantive and the second is institutional; neither could be opened by better argumentation.

| Documented failure | Gate that was near zero | Why that gate was closed | What N14 says would have been needed first |
|---|---|---|---|
| Intergovernmental panel on biotechnological risk | An expert consensus exists to be communicated | There is no uniform expert view on whether certain dangerous research should be done, or whether creating novel dangerous pathogens is a net harm | A pre-existing coalition around the problem large enough to represent a majority view |

| Documented failure | Gate that was near zero | Why that gate was closed | What N14 says would have been needed first |
|---|---|---|---|
| Parliamentary committee inquiry into unprecedented technological risk | The institution can do original research | Committees synthesise the views of witnesses and cannot perform response identification; launching the inquiry also required multiple sympathetic members to spend scarce time | Developed responses and prioritised policy proposals for the committee to synthesise |

Notice the asymmetry between the two. The panel's gate is closed because the underlying expert field disagrees; the committee's gate is closed by the institution's mandate. One is a fact about the world and the other is a fact about the vehicle. When an evaluator sees a proposal that uses an existing institution as the delivery vehicle, the mandate gate is the first thing to check, because it is cheap to check and it is usually the most decisive.

*Caution: the product form assumes the gates are independent. They are not. Work on response identification can build the coalition that raises political feasibility, and a visible coalition can raise opposition, which pulls feasibility down. The product is a diagnostic device for asking which constraint is binding, not a forecast model, and the independence assumption is the first thing to relax if numbers from the next subquestion are taken literally.*

**Step 2**

Turn the diagnostic into a computation once $gate_probabilities$ is supplied. No numerical gate inputs are provided in this example, and none are empirically calibrated to the corpus. The accompanying figure therefore shows the model's symbolic dependencies, not an invented gate profile or a current-policy estimate. The code below is a parameterized exercise: its arithmetic can illustrate the diagnosis in N14, but is not evidence that the chosen probabilities describe the world.

```
def chain_stats(gates):
    """Stage-gate diagnostics for a policy that must pass every
    gate."""
    V = 1.0
```

```
    for p in gates.values():
        V *= p

    order = sorted(gates.items(), key=lambda kv: kv[1])
    binding, p_binding = order[0]
    second, p_second = order[1]

    # dV/dp_i = product of the other gates = V / p_i
    marginal_gain_per_unit = {name: V / p for name, p in
    gates.items()}

    return {
        "chain_probability": V,
        "binding_gate": binding,
        "binding_gate_probability": p_binding,
        "second_gate": second,
        "second_gate_probability": p_second,
        "marginal_gain_per_unit": marginal_gain_per_unit,
        "ceiling_if_binding_perfected": V / p_binding,
        "ceiling_if_all_but_second_perfected": p_second,
        "is_binding_clearly_worst": p_binding * 3.0 < p_second,
    }


gates = dict(input_state["gate_probabilities"])
report = chain_stats(gates)

for key, value in report.items():
    output_state[key] = value

output_state["recommendation"] = (
    "repair the binding gate only if it is clearly the single worst;
    otherwise "
    "look for a policy or venue with no pair of small gates"
    if not report["is_binding_clearly_worst"]
    else "repair the binding gate"
)
```

Interpreting the output after numerical $gate_probabilities$ are supplied. The report then returns a single number for the whole policy, the name of the binding gate, the ceiling reachable by perfecting that gate alone, the ceiling reachable by perfecting everything except the second gate, and a flag for whether the binding gate is clearly the worst or merely the worst of a cluster. The flag is the operative output: it distinguishes a proposal with one problem from a proposal with several, and those call for different actions from a grantmaker.

A further warning about the marginal column. Because $\partial V/\partial p_i = V/p_i$, the marginal value of improving a gate is highest for the smallest gate. This makes the ranking look like a discovery, but it is a property of the functional form,

not a fact about the world. The genuinely informative comparison is between the marginal value of raising a gate and the marginal value of moving the same proposal to a different venue with a different gate vector, which the model as written cannot represent. ***Presenting the arithmetic as evidence rather than as arithmetic is exactly the failure mode to avoid in an interview.***

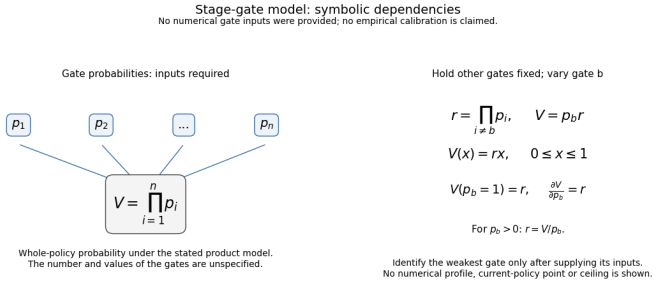


Figure 44: Graphical analysis for Step 2.

## Sub-Question 3: Windows, Entrepreneurs, Venues and Frames: The Four Levers That Move the Gates

N14 names Kingdon's multiple streams and policy windows, the policy entrepreneur, Schattschneider's expansion of conflict, Baumgartner and Jones on issue definition and venue shopping, and Schneider, Ingram and Stone on framing. Explain what each of these changes inside the decision condition from the first subquestion, and explain why coalition building is not monotone.

### Step 1

These five literatures are usually taught as rival accounts of the policy process. They are better read as levers acting on different terms of the same condition, and the mapping is the useful part.

Table 88: Five levers mapped onto the decision condition. Only the entrepreneur is under the funder's direct control; the others are conditions the funder can position around.

| Lever | What it actually changes | Who supplies it | Corpus anchor |
|---|---|---|---|
| **Policy window** | A time-limited shift in the politics stream. It changes the political-capital cost term, not the evidence term | Events: elections, crises, shifts in the governing coalition | N14 cites Kingdon and the story of North Atlantic current monitoring, where funding appeared after a scientist reached the prime minister |
| **Policy entrepreneur** | Couples problems, policies and politics. Supplies the labour that reduces the sponsor's private cost | An individual or organisation willing to hold the coupling across time | N14's own engagement is an act of entrepreneurship: the organisation identified problems and expected to produce responses |
| **Expansion of conflict** | Changes who counts in the audience. It alters the threshold and the composition of support and opposition at once | Whichever side the newly included audience favours | N14 cites Schattschneider's model and adds its own criticism, that visible coalitions can ignite opposition |
| **Venue shopping** | Replaces one gate vector with another, because each venue has its own threshold and its own veto points | Any actor with standing in more than one venue | N14 cites Baumgartner and Jones on issue definition and venue shopping |

| Lever | What it actually changes | Who supplies it | Corpus anchor |
|---|---|---|---|
| **Framing** | Determines which of several defensible descriptions of the same facts is the one being decided on | Advocates on both sides, and the media | N14 cites Schneider, Ingram and Stone and says framing decides which policies are implemented |

The first non-trivial reading. A policy window is a change in the politics term, and **a window is useless to an actor who has not already done the response-identification work**. N14 makes the point structurally when it says previously rejected policies can be implemented when a window opens, and then adds the mechanism, which is that a scientist with a long-standing proposal succeeded because policy analysts were already positioned to act on it. Kingdon's own framing supports the same reading: the three streams have to be coupled, and coupling takes preparation that cannot be produced on the day.

The second non-trivial reading follows. If coupling takes preparation, then the value of a policy grant is largely in the option value it creates, and option value is hard to see in the year it is granted. A grantmaker who only funds proposals that are already near a window will systematically underfund the preparedness on which windows are actually used. N14's own conclusion points the same way: it decided to spend more effort on building coalitions around the problem and identifying solutions, while continuing to respond reactively to opportunities.

*Both readings are inferences from a single organisation's account and should be offered as such. The corpus contains no base rate for how often windows are actually used, and N14's own caution is that the model can cause an actor to miss easy wins where the coalition required turns out to be surprisingly small.*

**Step 2**

Coalition building is where the model stops being a simple product. N14's own criticism is that building visible coalitions can ignite opposition, so an activity that raises support also raises the threshold. The minimal representation is a margin

$$M(v) = \big[S_0 + a\log(1+v)\big] - \Big[T_0 + S_{\text{opp}}(v)\Big]$$

where $v$ is visibility or organising intensity, $S_0$ is baseline support, $T_0$ is the baseline threshold, and opposition is assumed to mobilise more than proportionally once the issue becomes salient. The concavity of support and the convexity of

opposition reproduce N14's observation, and they are invented functional forms chosen to make the point, not estimated relationships.

```python
import numpy as np


def adoption_margin(v, base_support, support_gain, base_opp, opp_gain, threshold):
    """Linear-in-log support against quadratic opposition mobilisation."""
    support = base_support + support_gain * np.log1p(v)
    opposition = base_opp + opp_gain * np.square(v)
    return support - threshold - opposition


params = dict(input_state["coalition_parameters"])
visibility = np.linspace(0.0, 6.0, 601)
margin = adoption_margin(visibility, **params)
best = int(np.argmax(margin))

output_state["optimal_visibility"] = float(visibility[best])
output_state["margin_at_optimum"] = float(margin[best])
output_state["margin_at_zero_visibility"] = float(margin[0])
output_state["adoption_possible_at_some_visibility"] = bool(margin.max() > 0.0)
output_state["interior_optimum"] = bool(
    visibility[best] > 0.0 and visibility[best] < visibility[-1]
)
output_state["visibility_gap_to_adoption"] = float(-margin[best]) if margin[best] < 0 else 0.0
```

Once $coalition_parameters$ are supplied, the output is a pair of numbers that a grantmaker can act on: the visibility level that maximises the margin, and the gap to adoption at that level. If the optimum is interior, then more organising is a mistake beyond a point, and the grant should be sized to the optimum rather than to the ambition. If the gap is positive, no amount of organising by this coalition closes it, and the decision moves back to the second subquestion's test: is there a second closed gate, or a different venue. No numerical $coalition_parameters$ are supplied in this example. The accompanying diagram therefore shows the complete symbolic dependencies and diagnostic definitions; a numerical curve, adoption region, optimum and gap cannot be computed from the supplied material.

*Two honesty notes belong with this. First, the functional forms are invented; the substantive claim, that organising has diminishing returns while opposition can mobilise late and sharply, is an interpretation of N14's criticism and not something N14 measures. Second, the model treats opposition as a response to visibility, whereas in N14's examples opposition also exists independently, as with the blocked gestation-crate proposal where the special interest was al-*

*ready organised before the proposal was made.* **The distinction between pre-existing opposition and opposition created by the campaign is decision-relevant and the model as written cannot represent it.**

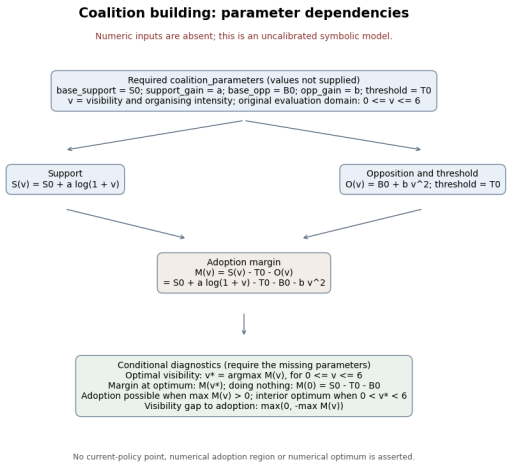


Figure 45: Graphical analysis for Step 2.

## Sub-Question 4: Venues Are the Real Decision Variable: The Three Governance Pathways and the Post-Enactment Chain

N06 maps AI governance as three pathways through seventeen activities. Treat each pathway as a venue with its own threshold, its own binding force and its own failure mode, and explain why the same policy idea has a different feasibility profile in each. Then trace a policy past enactment through regulation, standards, compliance, quality assurance and enforcement, and say where it leaks.

### Step 1

The staged model in the second subquestion tempts an evaluator into asking whether a proposal is feasible. N06 shows the question is malformed. It should be: feasible for whom, in which pathway, with which binding force? **Feasibility is a property of the policy-venue pair.**

Table 89: N06's three pathways read as three venues. The trade-off between them is not a preference ordering but a frontier: moving down the table buys feasibility and loses binding force.

| Attribute | Hard Governance | Industry-wide Self Governance | Company Self Governance |
|---|---|---|---|
| Who must be persuaded | Legislators, then the executive, then regulators and mandated standards bodies | A coalition of industry players capable of agreeing a norm or standard | An internal coalition of employees, investors or leaders, against the leadership's response |
| Binding force | Legal coercion, enforcement and verification, with sanctions such as fines and investigations | Reputation, procurement and market access, peer pressure, and sometimes regulator recognition of certification | Employment, investment and internal authority: mandates, orders, formal and informal sanctions |
| Typical speed and reach | Slowest to establish, broadest reach once enforced, and the only pathway that can compel non-signatories | Faster to agree, reaches participants and their supply chains, vulnerable to non-participants free-riding | Fastest and narrowest; binding only on one organisation and reversible by the next leadership decision |
| Characteristic failure | Dies in the political-feasibility gate, or is enacted and then diluted in regulation and guidance | Agrees a norm that is too weak to matter, or that competent but non-participating actors ignore | Loses the internal coalition when attention moves, and leaves no external trace |
| Funder lever | Lower treasury and political-capital cost by supplying research, drafting and coalition support | Fund the secretariat, standards work and coalition infrastructure; supply evidence the coalition can adopt | Support internal advocates and outsider pressure; smallest ticket sizes, hardest to attribute |

The non-trivial consequence: **a grantmaker choosing a venue is choosing which gate vector to accept, and the binding force of the venue is not a detail to be optimised later.** A voluntary industry norm and a statutory requirement aimed at the same behaviour are not the same instrument at different strengths; they act on different actors through different mechanisms, and the weaker one may be the only one that can be achieved in the decision window. Framed this way, venue selection is the operational version of the timeline-sensitivity question, because the value of a slow-binding instrument rises with the expected time before the capability arrives and falls with the probability of a short window.

*N06 also warns that its own map is incomplete, and the omissions are not random. Military and national security governance, supply chain and trade governance, multilateral soft governance, extralegal governance and academic and R&D governance are all left off. D04 lists international AI governance and law as eligible subject areas, and both live largely inside the omitted regions.* **An evaluator who reasons only from N06 will systematically underweight the areas where the map is thinnest, which is close to the opposite of good grantmaking.**

**Step 2**

Enactment is the middle of the chain, not the end. N06 traces it as: policy and lawmaking, then regulations and executive guidance, then standard-setting, then corporate compliance, then corporate governance and quality assurance, then AI development management, then enforcement and verification, then policy evaluation, whose results feed back into applied research and policy recommendations.

Two structural facts about that chain matter for an evaluator. First, the chain passes through the company's internal ecosystem, where the effective implementers are compliance, quality assurance and line management functions whose own incentives are the firm's, not the policy's. N06 shows these functions interpreting the rule, assessing the cost of compliance options, and adapting practice; the policy's intent is one input among several. So the feasibility question recurs at every handoff rather than being settled once. Second, N06 notes that policy evaluation is in some jurisdictions an extremely politicised tool, so the feedback loop that is supposed to correct the chain may itself be captured.

Table 90: The five macro-areas N06 leaves off its map, with the D04 subject area or proposal type through which a funder can still act. The mapping is the author's inference, not N06's or D04's claim.

| Omitted macro-area | Why it matters for catastrophic risk | How a funder might still reach it |
|---|---|---|
| Military and national security governance | Defence actors use AI under secrecy, which is the setting where catastrophic misuse and inadvertent escalation are least observable | D04 permits policy development and strategic analysis with a national-security framing; evaluate on the strength of access, not on public legibility |
| Supply chain and trade governance | Compute governance and export controls are the strongest available levers on frontier capability, and they act through commercial and customs instruments | D04 lists technical AI governance and names compute governance and secure, governable chips explicitly |
| Multilateral soft governance | Coordination among states is a precondition for avoiding race dynamics, even where instruments are non-binding | D04 lists international AI governance as an eligible subject area, including research into potential international institutions |
| Extralegal governance | Weight theft, cybercrime and proliferation outside the law are threats in themselves and shape the enforcement baseline | Most naturally a security-oriented part of technical AI governance, for example work on model-weight security and tamper resistance |
| Academic and R&D governance | It governs who is trained, what is published and which research agendas exist, which is upstream of everything else in the map | D04 funds training, mentorship and fellowship programmes as a distinct proposal type |

Finally, the funder's own gates. D04 assesses proposals on theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale, and

it states the scale preference in the first-time bracket of about $200,000 to $2 million per year over one to two years, with renewals after evidence of success. Two implications follow that are easy to miss. D04 says projects smaller than $200,000 per year should ideally be funded by other funders, which means part of the grantmaker's job is venue-shopping the applicant to a funder whose ticket band fits. And it lists strategic judgment as including reasoning transparency and scope sensitivity, which is a description of exactly the diagnostic behaviour this brief is building.

## Sub-Question 5: Worked Judgment Exercise: Diagnosing the Four-Pillar Biosecurity Plan by Venue

P06 reports a four-pillared plan to avoid engineered biological catastrophes: personal protective equipment such as elastomeric respirators, biohardening of buildings and infrastructure, pathogen-agnostic detection systems, and rapid medical countermeasures. Use the policy-process tools from this step to diagnose where the binding gate sits for this portfolio, produce two defensible diagnoses, and state what would change your mind. Do not claim any personal involvement in these proposals.

### Step 1

The first analytic move is one N14 supplies and that is easy to skip. The four pillars are **responses**, which is to say desired outcomes, not **policies**, which are methods of achieving outcomes. N14 makes exactly this distinction and says the community's problem was that it had not done the response and policy identification work. So the diagnostic question is not whether the four pillars are good. It is which policies could realise each pillar, in which venue, and what the gate profile of each policy-venue pair is.

Table 91: Each row separates the response from the policy and then names two venues. The two venue columns differ in almost every row, which is the finding: the lowest-threshold venue for a pillar is rarely the venue with binding force on the actors who matter.

| Pillar (response) | Plausible policy (method) | Venue with the lowest threshold | Venue with the strongest binding force | Gate most likely to be closed |
|---|---|---|---|---|
| Respiratory protection | Workplace respiratory-protection standards; employer duty-of-care requirements; procurement specifications for public buildings | Industry-wide standards and procurement rules | Occupational safety regulation in Hard Governance | Voter-group cost through employers, and enforcement capacity |
| Biohardening of buildings | Building-code amendments; ventilation and filtration requirements for new construction; insurance and liability incentives | Industry norms, codes and insurer practice | Building codes and standards bodies backed by law | Financial cost to builders and owners, concentrated on an identifiable group |
| Pathogen-agnostic detection | Public-health reporting mandates; wastewater and sentinel surveillance programmes; funding for detection R&D | Existing public-health agencies and research funders | Statutory reporting and surveillance obligations | Treasury cost, plus privacy and data-governance objections at the compliance stage |

| Pillar (response) | Plausible policy (method) | Venue with the lowest threshold | Venue with the strongest binding force | Gate most likely to be closed |
|---|---|---|---|---|
| Rapid medical countermeasures | Advance market commitments; push funding for platform technologies; emergency-use authorisation reform; clinical trial network readiness | Philanthropic and advance-purchase financing | Statutory authorisation and indemnity frameworks | |

Answer A, the binding-force answer. Fund policy-supply work in the Hard Governance venue for detection and countermeasures, accepting a long and uncertain timeline, on the grounds that the catastrophic scenario is one in which the actors who must act are precisely the ones who would not adopt a voluntary norm. The treasury-cost gate is the most tractable, and unlike the voter-cost gate it does not require the funder to compensate an identifiable bloc. The form of the grant is research, drafting, legal analysis and coalition support rather than advocacy for the merits, because the merits are not the binding constraint.

Answer B, the feasibility answer. For respiratory protection and biohardening, treat the binding constraint as procurement and private standard-setting rather than law, and fund the coalition and secretariat infrastructure that lets standards bodies and large buyers move, accepting narrower coverage and the risk of non-participants, because a norm that exists during the interval when law cannot be passed dominates a statute that does not exist. Under this answer the same dollars are cheaper per unit of behaviour change and produce evidence, including evidence of non-compliance, that is usable in the Hard Governance venue later.

Both answers are defensible, and the disagreement between them is not empirical. It is about whether the binding force of the venue is load-bearing for the specific pillar, which turns on the coverage question: do voluntary adopters include the actors whose behaviour determines the tail risk? That is the crux, and it is worth saying so explicitly rather than resolving it by preference.

**Step 2**

Make the crux explicit and testable. A crux is a possible fact that is uncertain, decision-relevant, and resolvable at reasonable cost; a risk is not a crux and an

unresolved value is not a crux either. The table below states the crux for each answer and the cheapest way to resolve it.

Table 92: Cruxes for the two diagnoses, with the cheapest resolution route. Selecting a crux is itself scope sensitivity: the third row is the crux a reviewer should insist on before either answer is funded.

| Answer | Crux: the fact that would flip it | Evidence that would resolve it | Approximate cost of resolution | |
|---|---|---|---|---|
| A: fund Hard Governance policy supply | That a same-session or near-term legislative window exists for detection or countermeasure authorities, which would make pre-positioning pay off soon | Structured interviews with legislative staff and committee calendars; historical base rate for comparable public-health authority legislation | Modest: weeks of an experienced policy analyst's time | Low |
| B: fund standards and procurement | That voluntary adopters include the actors whose behaviour determines the tail risk, or that procurement leverage reaches them | Adoption census by actor type and by category of facility; whether non-adopters are the high-consequence ones; whether buyers can condition contracts | Moderate: a scoped data-collection exercise over weeks | |

| Answer | Crux: the fact that would flip it | Evidence that would resolve it | Approximate cost of resolution | |
|---|---|---|---|---|
| Both | Whether the four pillars are complements or substitutes for each other, which determines whether funding one is a hedge or a partial investment | Descriptive analysis of which pillar fails first in a plausible release scenario; requires a threat model, not new data collection | Low to moderate, mostly analytic | |

A calibrated interview answer should do five things in order. Name the response-policy distinction and refuse to evaluate the pillars as if they were policies. Give the gate profile for the specific policy-venue pair rather than a verdict about feasibility. State the venue trade-off as a frontier between threshold and binding force. Name the crux and the cheap test. And state what is not claimed: no base rate exists in the corpus for how often windows open, the four-pillar list is a leadership statement in a CEO letter rather than a validated portfolio, and no source connects these pillars to a measured risk reduction.

*The closing caution is the one that separates a confident answer from an honest one. N14's own summary of its lesson is that initiatives which build on a majority view among experts are not useful unless a majority view already exists, and it notes that even within the organisation doing the work there was significant internal disagreement about what should be done.* **A four-pillar plan is a response list before it is a consensus, and the policy work that follows from it is legitimate precisely because the responses are not yet policies.**

# Build a crux table of disagreements within AI safety and EA philanthropy.

## Question

Across the corpus, which disagreements are **load-bearing** – that is, which premise could be flipped so that the recommended action reverses – and what evidence would move each side?

Deliverable: a **crux table**. Each row lists one disagreement, the stated positions with source attributions, the assumption each position depends on, and the observations that would move each side. The table must preserve uncertainty and must not merge one individual's view into their institution's voice.

## Background

Method. A **crux** is the earliest point of divergence whose reversal changes the conclusion – in D01 and L02 terms, the premise whose flip changes the sign or size of the warranted update. Two sources that reach the same conclusion from different premises do not disagree at the crux, and two sources that reach different conclusions from a shared premise diverge only downstream. This step localises the divergence instead of restating the conclusions.

Attribution ladder. Three levels are kept distinct: **institutional positions** (P01, P06, P08, D01, D04), **individual staff views** (P03 is one grantmaker's model; P07 is a set of personal suggestions; S09 and S11 are staff posts), and **third-party comparators** (P10 and P11-P12 from EA Funds and Manifund). Several sources are dated (P08 was published under Open Philanthropy) or self-interested (P11 and P12 argue for Manifund's own regranting model).

Constraint. The corpus contains genuine disagreement, including inside CG – P06 says staff disagree and P08 funds groups whose findings point in opposite directions. Where a source states a caveat that undercuts its own position, that caveat is a load-bearing part of the row rather than a footnote. P10's *"if you disagree with the above, that might be a good reason not to donate to us"* is the cleanest example.

## Sub-Question 1: Method: the crux test and the attribution ladder

Before any table, fix the inclusion rule. Three questions decide whether a disagreement earns a row: (1) is a disagreement stated or implied across attributable sources; (2) does flipping a premise reverse the recommended action, not merely adjust its magnitude; and (3) can the divergence be localised to a premise rather than to an entire worldview?

The third question does most of the work. A disagreement about values cannot be localised the way a disagreement about a fact can, and the table should say

so rather than paper over it. This is the same distinction the corpus uses when it separates empirical hedging, which spreads bets under factual uncertainty inside a shared value framework, from worldview diversification, which spreads allocations under uncertainty about the value framework itself.

**Step 1**

Define the crux formally, then fix the attribution ladder that every row must respect.

$$\text{Crux: a premise } p \text{ such that } \Delta V(p) = V_{\text{act}}(p) - V_{\text{omit}}(p) \text{ changes sign at some } p^*.$$

Table 93: Table M1. The attribution ladder. Every crux row carries one of these tags.

| Level | Definition | Corpus examples | How to read a disagreement here |
|---|---|---|---|
| **Institutional** | A published position of CG or a fund, binding on staff output. | P01, P06, P08, D01, D04, P05. | Treat as CG's stance, but check the date (P08 pre-rebrand). |
| **Individual staff** | A named person's model, suggestions or opinions. | P03 (Latham-Proenca), P07 (per-staff suggestions), S09 (Oehlsen), S11 (Austin Chen). | Never attribute to the institution; P07 explicitly says suggestions do not represent the institutional view. |
| **Third-party comparator** | An external fund or program with its own incentives. | P10 (EA Funds), P11 and P12 (Manifund). | Flag self-interest: P11 and P12 advocate a model Manifund operates. |
| **External primary** | Material cited by CG but authored elsewhere. | N06 (governance map), N14 (policy feasibility), N02 (competition modelling). | Useful for locating where CG's process argument sits against a wider literature. |

Two refinements. First, a disagreement between an institutional source and a labelled individual staff view is still a real disagreement; what changes is only who owns it. Second, a source disagreeing *with itself* (P06 names both the pull toward caution and the desire to resist it) is a tension worth a row, not an error to be smoothed.

## Sub-Question 2: The crux table

The core deliverable is split into two thematic tables so that each stays readable: Table C1 covers substantive AI-risk disagreements, and Table C2 covers philanthropy-process disagreements. Column five names the load-bearing assumption; column six gives, separately, what would move the A side and what would move the B side.

Read the two tables as the atlas's spine. Later steps reuse these rows; the cross-cutting assumptions in the next subquestion are exactly the assumptions that recur across multiple rows.

### Step 1

Table C1. Disagreements about the AI-risk object. Rows 1-5 are the cruxes that P03 makes explicit; rows 6-11 are openness, geopolitics, alarm and biorisk cruxes.

Table 94: Table C1. Substantive AI-risk cruxes. Rows 1-5 are the structure P03 lays out; rows 6-11 sit across openness, geopolitics, alarm and biorisk.

| # | **Crux (the question that divides)** | **Position A** | **Position B / C** | **Load-bearing assumption** | **What would move each side** |
|---|---|---|---|---|---|
| 1 | **Does transformative AI arrive soon enough that short-timeline work should dominate?** | P01 (Oehlsen) and P06 (CEO) report shortened timelines and treat fast progress as the default path; P01 has a team (Claire Zabel) planning for short timelines. | P08 records the 2016 estimate of a 10% chance of transformative AI by 2036 (and notes internal disagreement today); P08 cites a Metaculus median AGI date of June 2033; P03 gives mainstream economists 0.1-1.5%/yr; S11 (individual) says "timelines might be medium. or broad." | Whether grantmaker time and faster decisions should be reallocated toward short-timeline-robust work. | Move A: capability acceleration (METR time horizon, Epoch index) stalling. Move B: continued acceleration and further expert updating. |

| # | Crux (the question that divides) | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 2 | **Is a frontier AI growth takeoff real at the insiders' magnitude?** | 10-20%+ annual growth after full automation (Amodei; Korinek-Suh 2024); P03 uses $g_full$ = 18%. | Mainstream 0.1-1.5%/yr (Acemoglu 2025); FRI rapid-progress ~5% by 2050; IMF high-adoption ~4% cumulative over 10 years for sub-Saharan Africa (all cited in P03). | Whether catch-up growth is a top-tier cause area at all. | Move A: realised productivity data. Move B: observed automation of production. |
| 3 | **Does AI automate most physical tasks (the industrial explosion), or stall at cognitive tasks?** | P03 assumes robots automate most physical economic activities, from cooking to construction. | AI-skeptic objection in P03; the author concedes that without physical automation "the main point here becomes much less important." | Whether non-frontier economies can specialise in tradable physical activities. | Move A: robotics capability and cost evidence. Move B: physical-task automation staying marginal. |

| # | **Crux (the question that divides)** | **Position A** | **Position B / C** | **Load-bearing assumption** | **What would move each side** |
|---|---|---|---|---|---|
| 4 | **Is the adoption lag D a fixed shift, or do follower curves often fail to complete?** | P03 central model: the follower is the frontier curve shifted right by D, with central D = 10 years. | Rasmus Andersen's critique (quoted in P03): historical non-catch-up and follower curves that often never complete; the author concedes the same-endpoint assumption is "probably too optimistic." | Whether lowering D is a tractable philanthropic lever. | Move A: historical completion rates. Move B: infrastructure, energy, conflict and institutional evidence on differential endpoints. |

| # | **Crux (the question that divides)** | **Position A** | **Position B / C** | **Load-bearing assumption** | **What would move each side** |
|---|---|---|---|---|---|
| 5 | **After the detachment window, is the income gap a blip or permanent?** | P03 future 4 (convergence) makes the gap temporary; post-scarcity could make income differences irrelevant. | P03 future 1 (plateau) freezes the ~5x gap; future 3 (singularity) could lock in early advantage (Karnofsky lock-in). | Whether early-window catch-up deserves philanthropic attention. | Move A: evidence that post-window growth is idea-driven and non-rival. Move B: evidence that AI gains are embodied in capital with owners and price tags. |

| # | Crux (the question that divides) | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 6 | **Are open-weight models net good?** | S11 (individual, framed against a perceived hive-mind): "open weights models seem good"; "would go farther, Total Research Transparency, open data, full open source." | P08 lists information security measures to prevent unintended proliferation of dangerous AI systems, plus evaluations for dangerous capabilities such as cyberoffense and biological research. | Whether marginal misuse uplift from released weights exceeds diffusion benefits. | Move A: evidence that openness's marginal misuse risk is low. Move B: demonstrated misuse uplift from broad release. |

| # | Crux (the question that divides) | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 7 | **Transparency maximalism, or case-by-case privacy?** | S11: "secrecy bad, transparency maximalism." | P10: some impactful work requires privacy, and such grants are left unlisted but counted in aggregates; D01 values reasoning transparency while allowing that not all reasoning can be shared. | Whether disclosure of a given artefact increases risk. | Move A: case studies where secrecy caused harm. Move B: cases where disclosure caused harm. |

| # | Crux (the question that divides) | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 8 | **Is China primarily a rival to be outpaced, or a misunderstood actor?** | P08: "important for democracies to maintain their competitive advantage in AI over authoritarian countries" (institutional). | S11: "i don't understand why everyone's afraid of china" (individual). | Whether competitive framing is essential to risk reduction or is itself a risk factor. | Move A: evidence on authoritarian misuse of AI. Move B: evidence on misperception and the value of cooperation. |
| 9 | **Is fearmongering net harmful?** | S11: "general fearmongering seems bad." | P01: "the world isn't ready"; P06: "accelerating progress in AI only heightens the stakes"; P08 risk pathways (misuse, loss of control, institutional strain). | Whether alarm raises the right kind of attention from decision-makers. | Move A: persuasion evidence on backlash and credibility loss. Move B: evidence that under-alarm dominates over-alarm. |

| # | **Crux (the question that divides)** | **Position A** | **Position B / C** | **Load-bearing assumption** | **What would move each side** |
|---|---|---|---|---|---|
| 10 | **Are we in a pause, or in a pace-by-default world?** | S11: "we might be in a pause/pace-by-default world" (deceleration is plausible). | P06: fast progress "increasingly looks like the default path rather than an aberration"; P01 scaling urgency; P08 competition dynamics. | Whether to plan for deceleration versus acceleration. | Move A: an observed capability plateau. Move B: continued frontier progress. |

| # | Crux (the question that divides) | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 11 | **How severe is engineered and novel biological risk, and do the key studies agree?** | P06 highlights mirror-bacteria work and a four-pillar plan; P08 says AI could help bad actors engineer far more lethal and transmissible pathogens than those in nature. | P08 also funds a RAND paper finding that current language models do not substantially increase the risk of a biological attack, alongside a paper finding that models help novices – the two point in opposite directions, and both groups were funded. | Whether biorisk is a near-term priority and how large the marginal AI uplift is. | Move A: demonstrated uplift experiments. Move B: null or tightly bounded findings. |

Two rows are worth flagging as especially traceable to labelled individuals rather than to institutions: rows 6-8 and 10 rest on S11, an explicitly personal "disagreements with the AI safety hivemind" post. That does not weaken them; it changes what the interviewer should treat as CG policy.

**Step 2**

Table C2. Disagreements about how philanthropy itself should be structured. This is where third-party comparators (P10-P12) matter most, and where self-interest is a live confounder.

Table 95: Table C2. Philanthropy-process cruxes. Rows 3-7 come from one program (ACX) and are attributed to that program, not to CG.

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 1 | **Centralised professional grant-making, or decentralised regranting?** | P06 (institutional) defends CG's relatively light staffing ratio and argues that "moving grantmaking staff from our payroll to a grantee's is not necessarily a big efficiency win" – while agreeing regranting can be very impactful. | P12 (Manifund): "many people's first $100k of grants will beat a professional grantmaker's last $100k"; P11: regranting scales cheaply and seeds opportunities early. | Whether taste and dealflow scale better when dispersed than when centralised. | Move A: evidence of poor regrant-quality dispersion. Move B: comparative ROI (P11 cites a 53x stake in Timaeus). |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 2 | **Should money and taste be colocated or separated?** | P12 favours colocation (Leo Gao's microgrants, regranting), treats trust as the scaling bottleneck, and cites the "cursed game of philanthropy" principal-agent problem. | P12 also notes higher salaries buy better talent "but beware principal-agent problems"; P06 expects CG to make pay more competitive. | Whether trust scales with institutional size or decays as the community grows. | Move A: regrantor outcome data. Move B: evidence that centralised vetting outperforms dispersed judgment. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 3 | **Should philanthropic grants fund for-profit companies with charitable aims?** | P17: some of ACX's best past grants were in this category. | VCs already fund for-profits, so if VCs reject a company, be sceptical; the counterargument in P17 is that some companies clear the bar only once charitable effects are priced in. | Whether charity pricing changes the ranking of companies against the VC ecosystem. | Move A: examples that clear the charitable bar. Move B: evidence that VC rejection predicts unpromising charitable outcomes. |
| 4 | **Donate or invest when funding for-profits? (P17)** | Invest, so the program becomes self-sustaining without threatening nonprofit status. | Donate, to avoid formal nonprofit restructuring, legal work and new conflicts of interest. | Whether the legal and governance burden of a portfolio exceeds the reinvestment benefit. | Move A: feasible legal structures. Move B: evidence of conflict and backstab risk. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 5 | **Claim equity when a non-profit grantee becomes a for-profit? (P17)** | Ask for equity – funders before the transition otherwise get nothing while later funders get shares. | No action – legally dicey and unenforceable; P17 is explicitly uncomfortable making an unenforced gentleman's agreement. | Whether any instrument can bind a future, not-yet-existing company. | Move A: enforceable convertible instruments. Move B: practical non-recovery evidence. |
| 6 | **Nonfinancial (prestige or seal-of-approval) grants? (P17)** | Yes – several grantees said recognition and signal-boosting helped more than the cash. | No – at zero direct cost a cost-benefit analysis favours giving them to everyone, which destroys their signal value. | Whether the value of the signal is preserved or diluted by wider issuance. | Move A: measured signal value. Move B: evidence of dilution. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 7 | **How should retroactive impact-market grants be assessed? (P17)** | Wing it – ask the expert evaluators to value the outcome directly. | Blind the evaluators into judging the outcome as if it were a new charity at various funding levels. | Whether evaluation bias from knowing an outcome is a major or minor distortion. | Whichever design yields more accurate valuations. |
| 8 | **Is a large philanthropic funding wave coming, and should it be front-loaded?** | P10 (EA Funds): "we're betting that a lot of philanthropic money is about to arrive"; P10 cites CG's $825M GiveWell bump; P01 seeds the field now so organisations can absorb future funds. | P10's own caveat: "If you disagree with the above, that might be a good reason not to donate to us"; P01: CG prioritises impact over money moved and expects to keep saying no. | Whether organisations – not money – are the binding constraint on the field. | Move A: evidence that capital cannot be absorbed. Move B: evidence that organisations can absorb it. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 9 | **Marginalist equal-returns, or general-manager ownership?** | P06 (marginalist): equalise marginal returns within and sometimes across cause areas, so each marginal dollar produces roughly the same good. | P06 also names a shift, with scale, toward ambitious general-manager ownership (recruit founders, incubate organisations) and praises the approach; Project Tailwind and Resolution are examples. | Whether infra-marginal gains from ownership exceed the value of optimising the marginal dollar. | Move A: marginal-return data. Move B: infra-marginal gain examples (Project Tailwind, Resolution, MATS). |
| 10 | **Direct top charities, or novel system-changing charities?** | P16: AMF and GiveDirectly are "hard to beat." | P16: ACX is most interested in novel approaches to complex systems – technological breakthroughs, new institutions, targeted political change. | Whether novel approaches have comparable or higher expected value than proven direct charities. | Move A: evidence that novel approaches underperform. Move B: novel wins. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 11 | **How deep should grant due diligence go?** | P04: proportionate – sixteen-minute calls, one conversation plus reference checks for a small transition grant; the ERA AIxBio grant took under two months. | D01 offers GiveWell's AMF review (125 endnotes, archived sources) as a model, while conceding it is "more costly than it's worth for most analysts." | Whether expected error cost scales with grant size, reversibility and downside risk. | Move A: evidence that shallow reviews miss failure modes. Move B: evidence that depth rarely changes decisions. |
| 12 | **Preserve risk tolerance as CG scales, or accept increased caution?** | P06 and S08 want to resist the pull toward risk aversion and keep making weird-seeming bets on underappreciated topics. | P06 also names increased caution as "natural and somewhat appropriate" with more donors, larger budgets and more partnerships. | Whether partnership and reputational risk justify forgoing high-variance bets. | Move A: evidence that weird bets pay off. Move B: evidence of partnership or reputational cost. |

| # | Crux | Position A | Position B / C | Load-bearing assumption | What would move each side |
|---|---|---|---|---|---|
| 13 | **Is the binding constraint staff, or AI-enabled throughput?** | P01: CG is capacity-constrained and "the main thing preventing us from scaling our impact is staff time." | P06 and D02: AI enablement keeps the staffing ratio light (Grant Copilot, a BOTEC generator, subagent research teams). | Whether AI substitutes for grantmaker judgment or only for information collection. | Move A: productivity data. Move B: evidence that AI substitutes for judgment. |
| 14 | **How much should different types of career shift be valued?** | P05: shorter feedback loops – donation flows and career changes – are easier to evaluate, and the program risks being drawn to the evaluable ("are we being drawn too much toward what's easier to evaluate?"). | P05 also flags possible underinvestment in the longer-term, harder-to-measure impacts of shaping careers, and in media and narrative. | Whether evaluability correlates with impact. | Move A: long-run career-impact data. Move B: evidence that the evaluable is also the impactful. |

Notice that rows 1 and 13 interact: if AI substitutes for grantmaker time, the case for decentralised regranting weakens; if it does not, row 13's position A strengthens. The table keeps them as separate rows because they are separate cruxes, but the interaction is real.

## Sub-Question 3: Cross-cutting load-bearing assumptions and the evidence that would move each side

The same handful of assumptions recurs across rows. Isolating them shows where a single piece of evidence would resolve several rows at once – which is the real prize for an evaluator, because it is what makes a disagreement worth investigating rather than accepting.

Table C3 lists each recurring assumption, the rows that depend on it, its status in the corpus, and the evidence that would settle or shift it.

### Step 1

Table C3. Recurring assumptions, with the reach of each and the discriminating evidence.

Table 96: Table C3. Recurring assumptions across the crux rows, ordered by reach.

| **Load-bearing assumption** | **Rows that depend on it** | **Status in the corpus** | **Evidence that would settle or shift it** |
|---|---|---|---|
| Timelines are short enough that urgency dominates. | C1-1, C1-2, C1-4, C1-5, C1-10. | Contested inside CG (P01 and P06 update short; P08 records internal disagreement; S11 medium or broad); dated (P08's 2016 estimate). | Continued capability trends (METR horizon, Epoch index) versus an observed plateau. |
| Most physical tasks will automate. | C1-3, C1-5. | A P03 assumption, conceded as necessary for the main argument. | Robotics capability and cost curves. |

| Load-bearing assumption | Rows that depend on it | Status in the corpus | Evidence that would settle or shift it |
|---|---|---|---|
| Follower economies complete the same automation path. | C1-4. | Contested by Andersen in P03; the author concedes the assumption is probably too optimistic. | Historical completion rates and endpoint data. |
| Philanthropic capital can be absorbed by organisations. | C2-8. | A P10 bet; P01 seeds the field now; P06 names non-rival organisational capacity as a bottleneck. | Evidence on fundable capacity versus available capital. |
| Trust scales with institutional size. | C2-1, C2-2. | P12 argues it will not; P06 relies on staffing ratio and culture to keep it. | Regrantor and centralised-grant outcome data. |
| Grant dollars are additionality-relevant (no displacement). | C2-3, C2-9, C2-11. | GiveWell concedes in the AMF review that some AMF funding probably displaced other malaria funding and other uses (D01). | Counterfactual funding analysis: would another funder have filled the gap? |
| Disclosure of an artefact increases risk. | C1-6, C1-7, C1-8. | S11 rejects it; P08 and P10 accept it for specific artefacts and grants. | Case studies of disclosure harm. |
| Competitive framing is essential to risk reduction. | C1-8. | P08 institutional; S11 individual rejects it. | Evidence on authoritarian misuse versus cooperation gains. |

The figure below counts how many crux rows each assumption supports, which is a rough proxy for how much of the table a single piece of evidence would

move.

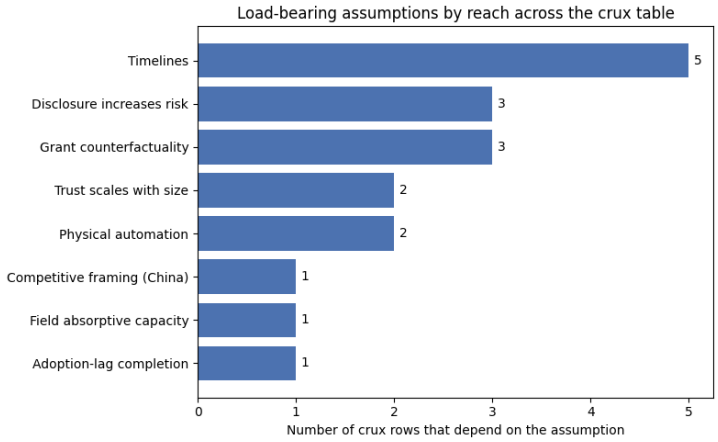


Figure 46: Graphical analysis for Step 1.

**Step 2**

The most consequential single crux is the intersection of C1-1 and C1-2: timelines multiplied by takeoff magnitude. P03 makes the structure explicit when it says everything in the post conditions on the insiders being roughly right, and it flags "all huge ifs" for AGI-pilled readers. That single conditional gate controls whether catch-up growth is the second most important problem in human welfare or a thought experiment.

Two other cruxes are nearly as consequential for a grantmaker's day-to-day, because they change the shape of a portfolio rather than its priority: C2-1, centralised versus decentralised grantmaking, and C2-11, due-diligence depth. These are the rows where an interview answer can show judgment without requiring a forecast about AGI.

One caveat the table must preserve: a crux is not the same as an unsettled question. Some rows (C1-6 through C1-8, all from S11) are stated as informal opinions rather than as positions with a documented evidentiary base, and the honest reading is that the corpus does not yet supply the evidence that would move either side. D01's discipline applies: say which claims are well-supported, which rest on a single person's read, and which have no support in the corpus at all.

## Sub-Question 4: Reading the table as an evaluator

The table answers a specific question and nothing more: where do attributable sources diverge, on what premise, and what would move each side. It does not

adjudicate. The point of the final subquestion is to state what the table licenses in an interview and where it would be misleading to lean on it.

**Step 1**

Use. Three rows do most of the work in a live answer. C1-1 and C1-2 let you show that you can hold a disagreement without resolving it: state both positions, name the assumption each depends on, and name the evidence that would move each side. C2-1 lets you take a position while conceding the strongest counter-argument, which is what the role brief calls comparing opposing arguments and identifying decisive considerations.

**Trap 1 – merging voices. The corpus's sharpest disagreements are between an institution and its own staff. Treating S11 as CG policy, or P03 as CG's position on growth, misstates the evidence. The ladder in Table M1 is the guard.**

**Trap 2 – confusing dated with current. P08 was published under Open Philanthropy and its editors note some content may be outdated; its 2016 timeline estimate is a historical anchor, not a present forecast. Where a source has a rebrand or a transition (P10 winding down LTFF and launching a new fund), the state change is itself relevant.**

**Trap 3 – treating a self-interested comparator as neutral. P11 and P12 argue for a model Manifund operates. That does not make the argument wrong; it means the argument carries an incentive weight that P06, defending CG's staffing ratio, also carries. The table flags both.**

**Trap 4 – resolving a crux by assertion. The role brief asks for calibrated, transparent reasoning. For a row like C1-8 (China), the corpus supplies an institutional position and one labelled individual disagreement and nothing else; the calibrated answer is to name the crux and decline to resolve it on this evidence, not to pick a side and sound confident.**

What the table does not settle. It cannot rank the cruxes by importance, because importance depends on the reader's prior about which premise is most uncertain – and that prior is itself contested. It also cannot tell you which side is right: the evidence that would move each side is, for most rows, not in the corpus. That is the calibration the role calls for, and stating it is stronger than pretending the table resolves anything.

# Map institutional differences among funders and organizations.

## Question

Map the institutional differences among the funders and organizations in the corpus: Coefficient Giving and Good Ventures, EA Funds' Transformative AI Fund, the LTFF legacy, Manifund, BlueDot, ACX Grants, SFF, Longview, Founders Pledge and others. Compare them on scale, speed, risk tolerance, transparency, grant size and grantees.

The deliverable is an institutional-difference map that preserves source attributions and distinguishes institutional positions from individual views and third-party comparators. The non-trivial jump is to treat funders not as points on six independent scales, but as different decision architectures whose observable attributes—scale, speed, risk posture, transparency, grant size and grantee access—are jointly determined.

## Background

The corpus contains first-party funder material, third-party comparators and individual views. Coefficient Giving (CG) is the central institution, described in P01, P04, P05, P06, P08, D01, D02, D04 and S01, S04, S07, S08 and S09. Good Ventures is CG's largest funding partner (P06, D04). EA Funds' Transformative AI Fund (TAF) is described in P10 and S02; the Long-Term Future Fund (LTFF) legacy is described in P10 and P11. Manifund regranting is described in P11 and P12. BlueDot Rapid Grants are described in P13, P14, D03 and S03. ACX Grants are described in P15, P16 and P17. SFF, Longview, Founders Pledge, AI Risk Mitigation Fund and Macroscopic appear mainly as named referents in D04, N02, P11 and S03.

The map must preserve uncertainty and institutional differences rather than merging organizations into one voice. For example, P10 is a fund launch with a fundraising interest; P11 is a Manifund fundraising post; P07 is explicitly a set of individual staff suggestions and not an institutional ranking; P08 was published under the former name Open Philanthropy and may be outdated; S10 and S11 are individual disagreements with parts of the AI safety community, not institutional positions of Manifund or CG.

## Sub-Question 1: The comparison schema: funders as decision architectures, not points on a scale

Build the schema that makes the institutional-difference map legible. The task names six axes: scale, speed, risk tolerance, transparency, grant size and grantees. The non-trivial move is to recognize that these axes are not independent and that ranking funders on each separately produces misleading conclusions. A funder's observable grant size band, decision speed and risk posture

are jointly chosen to implement a decision rule under a mandate and a capital constraint. The evaluator should compare architectures, not scores.

### Step 1

A funder can be represented as a mapping from opportunities to decisions. The mapping is parameterized by its mandate, capital, cost of evaluation per dollar, risk posture, transparency policy, grant-size band and dealflow mechanism. The non-trivial point is that the grant-size band is not exogenous: it is jointly determined with evaluation cost and risk posture so that the decision rule is implementable.

$$F_i : \mathcal{O} \to \{0,1\}, \qquad F_i(o) = 1 \iff \mathbb{E}[V_i(o)] \geq \tau_i, \qquad \tau_i = \tau(M_i, C_i, k_i, \rho_i)$$

Here $M_i$ is the mandate, $C_i$ is deployable capital per period, $k_i$ is the cost of evaluation per dollar, and $\rho_i$ is risk posture. Transparency and grantee access are separately observable but not separate objectives: they are constraints and search technologies that determine which opportunities reach the mapping at all.

```
from dataclasses import dataclass, field
from typing import Optional

@dataclass
class Funder:
    name: str
    mandate: str
    capital_scale: str
    eval_cost: str
    risk_posture: str
    transparency: str
    grant_size_band: str
    dealflow: str
    source: str

funders = [
    Funder(
        name="Coefficient Giving / Good Ventures",
        mandate="GCR: AI safety/security, biosecurity, global health, farm animal welfare, abundance and growth",
        capital_scale=">$1b directed in 2025; on track >$1b in 2026 for AI/biosecurity alone",
        eval_cost="high; shortened but still multi-week investigations for large grants",
        risk_posture="diversified; high-conviction bets; supports disagreeing experts; general-manager ownership",
        transparency="reasoning transparency standard (D01); publishes grants with delays",
```

```
        grant_size_band="$200k-$2m/yr preferred; flagship grants up to $212m",
        dealflow="proactive scouting, RFPs, referrals, incumbent grantee renewals",
        source="P01 P06 P08 D01 D04 S01 S04 S09"
    ),
    Funder(
        name="EA Funds Transformative AI Fund",
        mandate="technical AI safety, AI governance, field-building, forecasting; AI x bio/nuclear",
        capital_scale="seed ~$1m; hopes to raise several times more",
        eval_cost="low; full-time team, quarterly reports",
        risk_posture="tolerant of failure; power-law payoffs; early-stage",
        transparency="transparent by default; privacy exception for sensitive work",
        grant_size_band="typically <$150k; rare >$300k; not above ~$1m foreseeable",
        dealflow="open applications; proactive sourcing",
        source="P10"
    ),
    Funder(
        name="LTFF (legacy)",
        mandate="long-term future, including AI safety",
        capital_scale="820+ grants, just under $35m since 2017",
        eval_cost="low; small checks",
        risk_posture="high tolerance; first check to many orgs",
        transparency="public grants database",
        grant_size_band="small, speculative; average ~$43k",
        dealflow="open applications",
        source="P10 P11"
    ),
    Funder(
        name="Manifund regranting",
        mandate="AI safety regranting; broad early-stage",
        capital_scale="$2m in 2025; targeting $5m for 2026",
        eval_cost="low; part-time regrantors",
        risk_posture="high; hits-based; tolerant of failure",
        transparency="public dashboard; 5% operating fee",
        grant_size_band="pots of $100k-$200k for regrantors; individual grants from $2.5k",
        dealflow="regrantor networks; open applications to regrantors",
        source="P11 P12"
    ),
    Funder(
        name="BlueDot Rapid Grants",
        mandate="AI safety and biosecurity concrete work: compute, events, travel, tooling",
        capital_scale="77 grants / >$50k in a few months; scaling",
```

```
        eval_cost="very low; ~5 working days",
        risk_posture="high; small bets on people",
        transparency="public grantees list; no formal report",
        grant_size_band="$50-$10k",
        dealflow="community eligibility; open applications",
        source="P13 P14 D03"
    ),
    Funder(
        name="ACX Grants",
        mandate="novel charitable, scientific and system-changing projects",
        capital_scale="~$1m planned for 2025 round",
        eval_cost="moderate; volunteer expert evaluators",
        risk_posture="high; accepts unusual projects; now uses SAFEs/convertible grants",
        transparency="public results post with grant list and reasoning",
        grant_size_band="typically $5k-$50k; rare up to $100k",
        dealflow="ACX readers; open application",
        source="P15 P16 P17"
    ),
    Funder(
        name="SFF, Longview, Founders Pledge, AI Risk Mitigation Fund, Macroscopic",
        mandate="various; smaller and mid-scale AI safety and GCR opportunities",
        capital_scale="not described in corpus",
        eval_cost="not described",
        risk_posture="not described; SFF matched Manifund grant in N02",
        transparency="not described",
        grant_size_band="referred to by CG for projects below $200k/yr (D04)",
        dealflow="not described",
        source="D04 N02 P11 S03"
    ),
]

for f in funders:
    print(f"{f.name:42s} | {f.grant_size_band:38s} | {f.dealflow}")
```

Table 97: Table 1. Six dimensions and their common misreadings.

| **Dimension** | **What it measures** | **Common comparison error** | **Correction** |
|---|---|---|---|
| Scale | Capital available per period, not cumulative historical giving | Comparing total historical giving to annual giving | Normalize to the same period; distinguish committed from conditionally approved (P01 footnote) |
| Speed | Decision latency and time to money | Treating speed as independent of grant size | Speed is purchased by lowering evaluation cost per dollar and by tolerating error on small grants |
| Risk tolerance | Shape of the accepted payoff distribution | Equating risk tolerance with recklessness or with transparency | A funder can be risk-tolerant and opaque, or risk-averse and transparent |
| Transparency | Verifiability of reasoning and disclosure of grants | Treating all transparency as the same | Distinguish reasoning transparency (D01) from grant-list disclosure (P10) and privacy exceptions |
| Grant size | Typical and maximum grant | Comparing typical grant sizes across different stages | Stage matters: early-stage funders have low bands because their portfolio strategy requires many small bets |

| Dimension | What it measures | Common comparison error | Correction |
|---|---|---|---|
| Grantees | Eligible population and access mechanism | Treating the grantee list as revealed preference for cause areas | The access mechanism selects the grantee base; open applications, regrantor networks and proactive scouting find different opportunities |

**Step 2**

Derived arithmetic from the corpus reveals that the early-stage niche is not one band but several. The same label 'small grants' spans more than an order of magnitude, and the implied portfolio size differs accordingly.

```
ltff_total = 35_000_000  # P10: just under $35M
ltff_grants = 820        # P10: 820+ grants
manifund_2025 = 2_000_000  # P11: 13 regrantors gave away $2m
manifund_regrantors = 13
bluedot_total = 50_000   # P13: over $50,000
bluedot_grants = 77      # P13: 77 rapid grants
acx_planned = 1_000_000  # P16: ~$1m total
acx_grants_2025 = 42     # P15: 42 funded

print("LTFF average grant:", round(ltff_total / ltff_grants))
print("Manifund average pot:", round(manifund_2025 /
    manifund_regrantors))
print("BlueDot average grant:", round(bluedot_total / bluedot_grants))
print("ACX average grant (planned):", round(acx_planned /
    acx_grants_2025))
```

The arithmetic is not a ranking. It shows that BlueDot operates at sub-$1k average grant size, LTFF at roughly $43k, ACX at roughly $24k planned, and Manifund regrantor pots at roughly $154k. A comparison that puts these on one 'small grants' scale misses that they are different instruments for different portfolio designs.

## Sub-Question 2: Scale and speed: deployment architecture and the latency-size tradeoff

Compare the funders on scale and speed. The non-trivial jump is that speed is endogenous to grant size and to the cost of being wrong. A funder can be fast on small grants because the expected cost of a bad grant is small and the counterfactual is clear; large grants require more diligence because the grant itself moves the opportunity's marginal value and an error is costly. 'Fast' is therefore not a virtue in itself; a fast large grant may indicate under-diligence.

### Step 1

The corpus supports a structured comparison on scale, typical grant, decision speed and the mechanism of speed. The key distinction is between decision latency and time to money: BlueDot pays upfront by default (D03), while CG can take months even after a decision (P01 describes shortened investigations but not instant payment).

Table 98: Table 2. Scale, grant size and speed by institution.

| Institution | Scale | Typical grant band | Decision speed | Speed mechanism | Source |
|---|---|---|---|---|---|
| Coefficient Giving / Good Ventures | >$1b directed 2025; AI/biosecurity on track >$1b 2026 | $200k-$2m/yr preferred; up to $212m | days to months; $10m in 12 days example | shortened investigations, single additional sign-off, proactive scouting | P01 P06 D04 S01 |
| EA Funds Transformative AI Fund | seed ~$1m; hopes to raise several times more | <$150k typical; <$300k rare; <$1m | quick and reliable; target not yet specified | full-time team, open applications, pipeline | P10 |
| LTFF legacy | ~$35m total since 2017 | small speculative; avg ~$43k | fast small checks | small grants, low diligence cost | P10 P11 |

| Institution | Scale | Typical grant band | Decision speed | Speed mechanism | Source |
|---|---|---|---|---|---|
| Manifund regranting | $2m in 2025; target $5m 2026 | pots $100k-$200k; grants from $2.5k | fast, low-friction | part-time re-grantors, digital grant agreements, Stripe payouts | P11 P12 |
| BlueDot Rapid Grants | >$50k in a few months; scaling | $50-$10k | avg within a day; 9 in 10 within a week | community eligibility, simple application, upfront payment | P13 D03 |
| ACX Grants | ~$1m planned | $5k-$50k; rare $100k | annual round; October/November timeline | volunteer evaluators, Manifund admin | P15 P16 P17 |
| SFF, Longview, Founders Pledge, AI Risk Mitigation Fund | not described | referred for <$200k/yr | not described | not described | D04 N02 P11 S03 |

```python
ltff_years = 2024 - 2017 + 1
ltff_grants = 820
print("LTFF grants per year:", round(ltff_grants / ltff_years))

bluedot_months = 3
bluedot_grants = 77
print("BlueDot grants per year (annualized):", round(bluedot_grants *
    12 / bluedot_months))

acx_rounds_per_year = 1
acx_grants = 42
print("ACX grants per round (annual if one round):", acx_grants)
```

The throughput arithmetic shows why the bands differ. LTFF averaged roughly

117 grants per year over its life; BlueDot annualizes to roughly 308 grants per year; ACX funds 42 in a single annual round. Higher throughput is possible only when the grant size and evaluation cost per grant are low. This is the latency-size tradeoff in operational form.

## Sub-Question 3: Risk tolerance and transparency: two axes that are not the same

Distinguish risk tolerance from transparency. The non-trivial jump is that risk tolerance is about the shape of the payoff distribution a funder accepts, while transparency is about the verifiability of reasoning and disclosure of decisions. A funder can be risk-tolerant and opaque (some regrantors) or risk-averse and transparent (GiveWell-style). Treating them as one axis collapses a real institutional difference.

### Step 1

A power-law portfolio model makes the link between risk posture and grant size explicit. If each grant independently has probability p of a hit returning R and 1-p of returning zero, expected value per grant is pR. A portfolio of n grants has probability 1-(1-p)^n of at least one hit. For a target probability q of at least one hit, the required number of grants is n = ln(1-q)/ln(1-p). With a fixed budget B, average grant size is B/n. Rare hits force small average grants or a lower probability of any hit.

$$n = \frac{\ln(1-q)}{\ln(1-p)}, \qquad \bar{g} = \frac{B}{n} = \frac{B\ln(1-p)}{\ln(1-q)}$$

```python
import math

def grants_needed(p, q=0.9):
    return math.ceil(math.log(1 - q) / math.log(1 - p))

budget = 2_000_000
for p in [0.05, 0.02, 0.01, 0.005]:
    n = grants_needed(p)
    print(f"p={p:.3f}: n={n}, avg grant=${budget / n:,.0f}")
```

At p=0.01 and a $2m budget, a 90 percent chance of at least one hit requires about 230 grants and an average grant of about $8.7k. This matches the BlueDot and Manifund small-grant niche. A funder that believes hits are more common, or that accepts a lower probability of any hit, can write larger checks. The risk posture and the grant-size band are therefore jointly determined, not independent policy choices.

Table 99: Table 3. Risk posture and transparency mechanisms are different axes.

| **Funder** | **Stated risk posture** | **Instrument** | **Transparency mechanism** | **Relation to reasoning trans-parency** |
|---|---|---|---|---|
| Transformative AI Fund | tolerant of failure; power-law payoffs | grants | transparent by default; quarterly reports; privacy exception | disclosure of grants, not a reasoning standard |
| Manifund | hits-based; tolerant of failure | grants via regrantors | public dashboard; asks donors for 5% operating fee | operational and financial trans-parency, not standardized reasoning transparency |
| BlueDot Rapid Grants | small bets on people and projects | grants; sometimes reimburse-ment | public grantee list; no formal report | minimal reporting; not D01 reasoning transparency |
| ACX Grants | accepts unusual projects; now uses SAFEs/con-vertible grants | grants, SAFEs, convertible grants | public results post with grant list and reasoning | closer to reasoning transparency but not standardized |
| Coefficient Giving | diversified; high-conviction bets; supports disagreeing experts | grants | D01 reasoning transparency standard | the standard itself; not always met in every post |

## Sub-Question 4: Grantee access and dealflow: the funnel, not the list

Map grantee access and dealflow. The non-trivial jump is that the grantee base is selected by the access mechanism, and funders are nodes in a directed

pipeline from seed to scale. The same organization can appear at multiple stages: Manifund regranted $143k to Timaeus, which later merged with Resolution, which received CG's $160m grant (P11). S03 reports that Incubator Week participants raised pre-seeds from CG, Longview and BlueDot. The map should be a funnel, not a league table.

**Step 1**

The observed and claimed funding pathways form a directed graph. The graphical analysis below is a stylized representation of edges named in the corpus; it is not a complete map of the ecosystem.

Table 100: Table 4. Access mechanisms and observed or claimed next-stage pathways.

| **Funder or intermediary** | **Access mechanism** | **Eligible population** | **Observed pathway** |
|---|---|---|---|
| Coefficient Giving / Good Ventures | proactive scouting, RFPs, referrals, renewals | established orgs, founders via Project Tailwind, large projects | terminal large-scale funder; seeds via Tailwind (S09) |
| Transformative AI Fund | open applications, proactive sourcing | anyone starting something new in transformative AI | early-stage; may hand off to CG, Longview, SFF |
| Manifund | regrantor networks, open to regrantors | part-time grantmakers with taste and dealflow | regrants to Timaeus, Apollo, ChinaTalk; some later raised from CG (P11) |
| BlueDot Rapid Grants | community eligibility; open applications | BlueDot course participants, alumni, facilitators, active community | funds pre-seeds; incubator participants raised from CG, Longview, BlueDot (S03) |
| ACX Grants | open application to ACX readers | ACX readers with novel projects | sometimes shops projects to VCs and philanthropists (P16, P17) |

| Funder or intermediary | Access mechanism | Eligible population | Observed pathway |
|---|---|---|---|
| SFF, Longview, Founders Pledge, AI Risk Mitigation Fund | not fully described; SFF matches Manifund grants | smaller and mid-scale projects | receive referrals from CG D04; SFF matched Manifund grant N02 |
| Incubator Week (BlueDot) | application and recruitment | founders pivoting to AI safety | pre-seeds from CG, Longview, BlueDot (S03) |

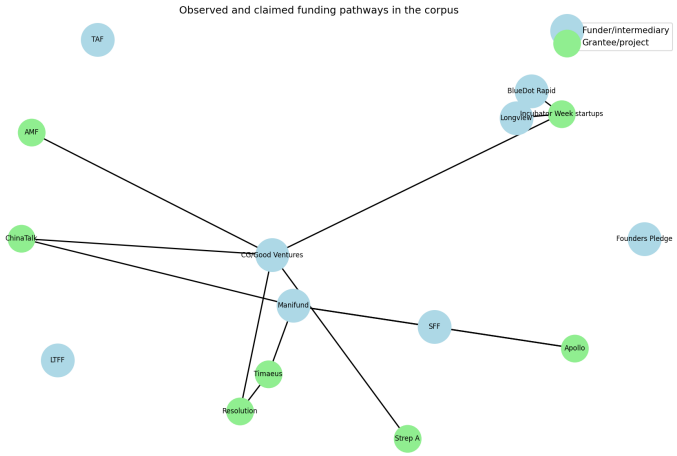


Figure 47: Graphical analysis for Step 1.

## Sub-Question 5: Institutional versus individual versus third-party: weighting the sources

Preserve the epistemic status of each claim. The non-trivial jump is that the same institution's name can appear in sources with very different evidential weight. P07 is a set of individual staff suggestions and is explicitly not an institutional ranking. P06 is the CEO's personal reflection but is institutionally authoritative. P08 was published under the former name Open Philanthropy and may be outdated. P10 is a fund launch with a fundraising interest. P11 is Manifund's fundraising post. P15 and P16 are third-party comparators. S10 and S11 are individual disagreements with parts of the AI safety community, not institutional positions.

**Step 1**

The table below assigns a provenance status and a calibration note to each source family. The evaluator should prefer the source that is closest to the decision and least distorted by incentives, while still using self-interested sources for operational facts they assert about themselves.

Table 101: Table 5. Source provenance and calibration for funder claims.

| Source | Institution or author | Status | Key claim about funders | Calibration note |
|---|---|---|---|---|
| P06 | Coefficient Giving CEO | institutional but personal in tone | marginalist shift, general-manager ownership, scaling | authoritative for CG direction; not a formal policy document |
| P07 | CG staff | individual views, not institutional | individual giving suggestions | explicitly not an institutional ranking; use for staff judgment, not CG policy |
| P08 | CG (as Open Philanthropy) | institutional but dated | three pillars: visibility, safeguards, capacity | may be outdated; note the editor's note |
| P10 | EA Funds | institutional fund launch; self-interested | TAF scale, speed, transparency, grant bands | operational commitments are usable; fundraising context should be discounted |
| P11 | Manifund | institutional; self-interested | regranting structure, part-time grantmakers, examples | examples are selected hits; useful for mechanism, not for base rates |

| Source | Institution or author | Status | Key claim about funders | Calibration note |
|---|---|---|---|---|
| P15, P16, P17 | ACX Grants / Scott Alexander | third-party comparator | ACX scale, process, SAFEs, evaluator roles | individual project, not a formal institution; process details are usable |
| S10, S11 | Austin Chen | individual views | disagreements with AI safety hivemind on openness, China, politics, timelines | not a Manifund institutional position; use as evidence of internal disagreement |
| D04 | Coefficient Giving RFP | institutional and operational | CG grant size preference; referrals to SFF, LTFF, Founders Pledge, AI Risk Mitigation Fund, Longview | updated Sept 2026; use for operational fit |

**Step 2**

The funder map also contains genuine disagreements about strategy, not only differences in scale. Preserving them is part of the evaluator's job.

- Centralised large funder versus dispersed regrantors: P06 and P01 describe CG scaling with in-house grantmakers and RFPs; P11 argues that part-time regrantors can find opportunities before centralised funders and that regranting is more scalable. The crux is whether dealflow or evaluation capacity is the binding constraint.
- Early-stage speed versus large-scale diligence: P10 commits TAF to fast, small, failure-tolerant grants; D04 says CG prefers $200k-$2m per year and refers smaller projects to other funders. The crux is whether the same opportunity is better served by a fast small grant or a slower larger grant, which depends on the stage and the counterfactual.

- Openness and China: S10 and S11 express individual disagreement with parts of the AI safety community on open weights, transparency, China and politics. These are not institutional positions of Manifund or CG, but they mark live disagreements an interviewer may probe.
- Transparency as disclosure versus transparency as reasoning: P10's transparent-by-default policy is about grant disclosure; D01's reasoning transparency is about confidence, support and key considerations. Both are valuable, but they are different standards and should not be conflated in an interview.

## Sub-Question 6: Worked exercise: matching two hypothetical opportunities to funders

The knowledge summary proposes an exercise: for two hypothetical opportunities, choose the best-fit funder and justify the mismatch risks. The non-trivial jump is that funder fit is a conjunction of grant size band, access eligibility, mandate alignment, stage and timeline. The most common mismatch is not wrong cause area but right cause area with the wrong instrument or stage.

### Step 1

Opportunity A is invented for this exercise: a two-person team with a working prototype and a $60k request for a six-month AI governance research pilot. It has no institutional home and one team member is a BlueDot alum. Opportunity B is also invented: an established 501(c)(3) with a three-year track record and a $1.5m per year request over two years to scale a safety evaluation program.

Table 102: Table 6. Hypothetical opportunities for the funder-fit exercise.

| Attribute | Opportunity A | Opportunity B |
|---|---|---|
| Stage | prototype | established, scaling |
| Request | $60k for 6 months | $1.5m/yr for 2 years |
| Institutional home | none | 501(c)(3) |
| Cause | AI governance research | AI safety evaluations |
| Team | 2 people, one BlueDot alum | 15 people, 3-year track record |

```
funders = [
    {"name": "BlueDot Rapid Grants", "min": 50, "max": 10_000,
    "eligibility": "community", "stages": ["prototype", "early"]},
    {"name": "Manifund regranting", "min": 2_500, "max": 300_000,
    "eligibility": "open", "stages": ["prototype", "early"]},
    {"name": "Transformative AI Fund", "min": 25_000, "max": 300_000,
    "eligibility": "open", "stages": ["idea", "prototype", "early"]},
```

```
    {"name": "Coefficient Giving", "min": 200_000, "max": 2_000_000, "eligibility": "open", "stages": ["established", "scaling"]},
    {"name": "Longview / SFF / Founders Pledge", "min": 50_000, "max": 2_000_000, "eligibility": "open", "stages": ["prototype", "established", "scaling"]},
]

opportunity_a = {"name": "A", "grant": 60_000, "stage": "prototype", "eligibility": "community"}
opportunity_b = {"name": "B", "grant": 1_500_000, "stage": "scaling", "eligibility": "open"}

def match(o):
    print(f"Opportunity {o['name']}")
    for f in funders:
        reasons = []
        if o["grant"] < f["min"]:
            reasons.append(f"below min ${f['min']:,}")
        if o["grant"] > f["max"]:
            reasons.append(f"above max ${f['max']:,}")
        if o["stage"] not in f["stages"]:
            reasons.append(f"stage {o['stage']} not typical")
        if f["eligibility"] == "community" and o["eligibility"] != "community":
            reasons.append("not community-eligible")
        if not reasons:
            print(f"  FIT: {f['name']}")
        else:
            print(f"  MISMATCH: {f['name']} - {', '.join(reasons)}")
    print()

match(opportunity_a)
match(opportunity_b)
```

Table 103: Table 7. Funder-fit recommendations and mismatch risks for the hypothetical opportunities.

| **Opportunity** | **Best-fit funders** | **Mismatch risks** |
|---|---|---|
| A | Transformative AI Fund (<$150k); Manifund regranting; BlueDot Rapid Grants if staged to ≤$10k and community-eligible | CG: below $200k-$2m/yr preference and likely referred elsewhere (D04). BlueDot: $60k exceeds the $10k ceiling; could fund a smaller scoped component. No institutional home may require fiscal sponsorship for some funders. |
| B | Coefficient Giving / Good Ventures ($200k-$2m/yr); Longview, SFF, Founders Pledge (mid-scale) | TAF: above its typical <$150k, rare >$300k and <$1m ceiling; stage mismatch. BlueDot: far above $10k. Manifund: above regrantor pot sizes. The evaluator should not send a scaling request to an early-stage funder just because the cause matches. |

The exercise teaches three evaluator habits. First, check the grant size band before the cause area, because a mismatch on size can waste the applicant's time and the grantmaker's. Second, check the stage: early-stage funders and scale funders are not substitutes. Third, check the access mechanism: an opportunity that is not community-eligible cannot use BlueDot's speed, and an individual without a fiscal sponsor may need a different vehicle.

# Create an uncertainty register for major claims.

## Question

Build an uncertainty register for the major claims in the corpus. For each claim about **timelines, takeoff speed, AI risk, biorisk, intervention effectiveness, ROI and catch-up growth**, record who holds it, how confident they say they are, what kind of support they think they have for it, what would move it, and who disagrees. The register is an instrument, not an answer sheet: its job is to make the non-trivial jumps in logic and interpretation visible rather than to resolve them.

Three distinctions must survive intact through every row: **institutional position** versus **named individual staff view** versus **third-party comparator**; source-supported claim versus inference versus proposed exercise; and a published number versus a hedge that was never quantified. *Converting a qualitative hedge into a probability the source did not publish is the single most damaging error this register can make.*

## Background

The register is deliberately modelled on **D01 (Reasoning Transparency)**, which asks a reader of any analysis: how should I update in response to this? D01 answers with three cheap moves — open with linked key takeaways, mark which considerations are load-bearing, and indicate confidence and kind of support for each major claim. It also offers an explicit taxonomy of support kinds (from careful examination of studies you feel qualified to assess, down to a vague impression based on reading or a general intuition about how the world works) and a set of confidence encodings, from the word 'plausible' through 'seems likely' (>50%) to a 70% confidence interval. The register takes those two vocabularies and makes them mandatory fields.

D01 also warns that words travel badly: *'plausible' can be interpreted differently by different readers*, and GiveWell-style documents pay for that precision in effort. The corpus we are registering is full of exactly this ambiguity: 'I am uncertain about these projections' (P01), 'timelines might be medium. or broad.' (S11), 'it's too hard to think about' (P03 on the singularity). A register that silently assigns numbers to these hedges would be worse than no register.

Second, the corpus is not one voice and should not be flattened into one. P08 says explicitly that Coefficient aims to support a wide range of experts, including those who disagree with CG and each other, and it gives the worked counterexample of two CG-funded papers that point in opposite directions on AI's impact on biorisk. P06 records the CEO changing her own prior about AI timelines, and enumerates internal tensions (marginalist optimisation versus general-manager ownership; scaling versus agility; near-term AI giving versus longer-term absorptive capacity). S11 is one person's sketch of where they disagree with the AI-safety consensus; S02, S05, S06 and S09 are individual staff announcements

on LinkedIn and X, not institutional documents. N14 is a reprint of an older first-person process account from the Centre for Effective Altruism. N02, N03 and N04 are third-party Manifund or forum applications.

Third, several corpus items are dated, partial or self-interested, and those properties belong in the register as fields, not as vaguely remembered caveats. P08 was published under the former name Open Philanthropy and carries an editor's note that some content may be outdated; the 2016 '10% by 2036' estimate is ten years old in a field where the authors themselves say their views have moved. P03 is one grantmaker's model with an explicit coda of objections. P02 calls its ~5,500x modelled social return early and uncertain. N02 is a fundraising page arguing for its own neglectedness. P07 is a list of individual staff suggestions explicitly not representing the institutional view.

The purpose of the exercise is interview-grade. An evaluator who is asked 'how likely is transformative AI?' should be able to say whose number they are quoting, what kind of support it rests on, and what would move it — and equally able to say 'no number exists for that claim in the corpus, here is the hedge, and here is the evidence that would license a number.' That skill is the difference between reasoning transparency and confident noise.

## Sub-Question 1: Subquestion 1: Register architecture, confidence vocabularies and tagging discipline

Design the register before populating it. What fields does a row need so that **locus, confidence, kind of support, crux and provenance** are all recoverable without re-reading the source? Which confidence vocabularies does the corpus itself license, and which are we importing from D01 and D02?

A row should be able to answer four questions about any claim: **who says it, with how much confidence, on what kind of support, and what would change it**. Fields that do not serve one of those four functions are clutter and should be cut.

### Step 1

The row carries ten fields. The three load-bearing ones are **locus** (who is speaking), **confidence** (how strongly, in the source's own encoding) and $support_k ind$ (on what basis).

Table 104: Table 1. Register schema. The disagreement field is the one most often degraded into a vague gesture; it must name the clause of disagreement.

| Field | What it holds | Which of the four questions it serves |
|---|---|---|
| `id` | Stable short key, e.g. T1, K2, R4, B3, E5, G2. Facet initial plus a number. | Retrieval only. |
| `facet` | One of: timelines, $takeoff_speed$, $ai_risk$, biorisk, $intervention_effectiveness$, roi, $catch_up_growth$. | Grouping; makes missing facets visible. |
| `claim` | One sentence, stated so that it could in principle be false. | Who says it. |
| `locus_kind` | institution, $named_individual_staff$, or $third_party_comparator$. | Who says it — and at what authority level. |
| `locus` | Named holder and source pointer, e.g. 'Emily Oehlsen (P01)', 'CG (P08)', 'Austin Chen (S11)'. | Who says it, traceably. |
| `confidence_kind` | probability, interval, $word_numeric_anchored$, or $qualitative_only$. | How strongly, and in what encoding. |
| `confidence_value` | The source's own words or numbers. Never a number the source did not publish. | How strongly. |
| `support_kind` | One or more labels from the D01 taxonomy in the next table. | On what basis. |
| `crux` | The specific observation, datum or argument that would move the claim. Phrase it as a test, not a topic. | What would change it. |

| Field | What it holds | Which of the four questions it serves |
|---|---|---|
| `disagreement` | Who holds a non-identical position, and on which clause they differ (not merely 'they disagree'). | Who would change it — and where the crux actually sits. |
| `status_tag` | $source_supported$, inference, or $proposed_exercise$. | Provenance discipline. |

D01 licenses four encodings of confidence, and the corpus uses all four. The point of recording the encoding rather than a normalised number is that a hedge is not a probability.

Table 105: Table 2. Confidence encodings drawn from D01. Register convention: keep the source's encoding; never silently upgrade a word to a number.

| D01 example | Encoding | What it signals about effort |
|---|---|---|
| Transformative AI within 20 years | probability with a robustness qualifier: 'at least 10% with moderate robustness, at least 1% with high robustness' | Thought hard; the qualifier encodes that the number is not robust to worldview shifts. |
| Years to high-level machine intelligence | interval: 'a 70% confidence interval ... 10-120 years, though that estimate is unstable and uncertain' | Worth a probabilistic statement at the conclusion, plus a stability caveat and a footnote defining 'unstable'. |
| Regular high-quality sleep is important | word: 'widely believed, and seems likely' | Peripheral claim, no investigation; 'seems likely' means >50% and no more. |

| D01 example | Encoding | What it signals about effort |
|---|---|---|
| CBT-I is the most common behavioural insomnia treatment | word: 'plausibly' | 1-3 hours of research could have raised confidence; not worth it for a non-central claim. Note D01's warning that 'plausible' reads differently to different readers. |
| Completeness of a literature search | precise probability in a footnote: 'I'm 70% confident there are fewer than 5 ... which I did not find' | Colloquial main text plus a precise, checkable footnote. This is the pattern the register should imitate. |

The support-kind vocabulary is D01's, lightly edited. Its value is diagnostic: two claims with the same confidence and different support kinds call for entirely different due diligence.

- **Strongest:** $own_{d}etailed_{a}nalysis$; $careful_{e}xam_{q}ualified$; $verifiable_{f}act$ with a source; $follows_{f}rom_{s}upported$.
- **Middle:** $careful_{e}xam_{w}eakly_{q}ualified$; $shallow_{s}kim_{q}ualified$; $expert_{o}pinion_{a}ssessable$; $simple_{a}rgument_{r}obust$; $complex_{a}rgument_{s}trong$.
- **Weakest, and the ones that must be labelled as such:** $vague_{i}mpression$; $general_{i}ntuition$; $expert_{o}pinion_{n}ot_{a}ssessable$; $remembered_{u}nverified$; $simple_{a}rgument_{q}uestionable$.
- A combination label is legitimate and common; D01 lists 'a combination of any of the above' explicitly, and warns against letting an unassessable expert opinion borrow the status of an assessable one.

**Step 2**

The register is only as good as its discipline, and the failure modes are predictable. Each anti-pattern below has a specific fix encoded in the schema.

- **Numeric laundering.** Writing 'roughly 15%' next to a claim whose source said 'plausible'. Fix: $confidence_{v}alue$ holds the source's own string; any number we supply is a separate inference row.
- **Voice merging.** Attributing P03's catch-up model or P06's personal timeline update to Coefficient as an institution. Fix: $locus_{k}ind$ and locus are mandatory and non-empty.
- **Disagreement as gesture.** 'Some disagree' tells an interviewer nothing. Fix: the disagreement field names the clause and the counter-holder, e.g.

'S11 disputes the China-fear clause, not the capability clause'.

- **Crux as topic.** 'More research on takeoff' is not a crux. Fix: crux must name an observable that would resolve it, in the spirit of P04's rule — ask early and often what piece of evidence would be most likely to change my mind.
- **Status flattening.** Letting an inference or an exercise ride on the credibility of a source-supported claim. Fix: $status_tag$ on every row, and inferences that depend on source rows carry their own dependency list.

## Sub-Question 2: Subquestion 2: Timelines and takeoff-speed entries

Populate the register with every timeline and takeoff-speed claim in the corpus, keeping institutional estimates, named staff views and third-party characterisations in separate rows. For each, record the confidence encoding the source actually used and the kind of support they think they have.

The interesting content is not the numbers but the **shape of their disagreement**: which claims are about the timeline itself, which are about the speed of the transition once it starts, and which are about the growth rate after full automation. Those are three different objects and the corpus sometimes describes them with the same words.

A metric-doubling claim also has a numeric handle that can be turned into a decision-relevant date: if the length of a software task that a frontier agent can complete 50% of the time doubles every $d$ months, then the time to reach a horizon $H$ from a current horizon $h_0$ is $t = d \log_2(H/h_0)$. That is a computation the register should carry rather than gesture at.

### Step 1

Timeline entries. Note how many of them are $qualitative_only$: the corpus is unusually willing to say 'I am uncertain' without supplying a number.

Table 106: Table 3. Timeline entries. Six of seven claims are not numbers, and the single institutional number is ten years old.

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement |
|---|---|---|---|---|---|---|
| T1 | Transformative AI may arrive within the next few years; 'a lot of impactful work urgently needs doing'. | Named staff, Emily Oehlsen (P01) | Qualitative only; explicitly 'I'm uncertain about these projections'. | $vague_impression$ + $general_intuition$ (capability progress, agentic coding, revenue growth, personal use) | Capability benchmark trajectories and whether agentic-coding gains persist. | P03 conditions on insiders versus mainstream economists; P06 flags bubble dynamics. |
| T2 | Short-timeline predictors have been more accurate than the CEO expected; continued fast progress looks like the default path rather than an aberration. | Named staff, CEO reflection (P06) | Directional update only; 'I have been struck by how much more accurate'. | $verifiable_fact$ (Epoch Capabilities Index, METR time horizon) + $vague_impression$ | Whether measured progress continues or plateaus after 2026. | Self-caveated: 'this is exactly the sort of extrapolative expectations dynamic that can drive bubbles'. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement |
|---|---|---|---|---|---|---|
| T3 | At least a 10% chance of transformative AI by 2036 (2016 estimate). | Institution, CG/OP (P08) | Probability with a robustness qualifier; D01's parallel phrasing is 'at least 10% with moderate robustness'. | *simple*$_a$ + *own*$_d$ | Whether the base rate moved; P08 says only that many staff now say 'considerably higher' by 2035. | The unquantified 2025 update is weaker evidence than the 2016 number; do not treat the update as a revised probability. |
| T4 | AI insiders predict sustained 10-20%+ annual growth after full automation; mainstream economists put AI-induced excess growth at 0.1-1.5%/yr. | Third party characterisation of others' views (P03, citing Amodei and Acemoglu 2025) | Two disputing camps with very different numbers; P03 conditions on the insiders being roughly right. | *expert*$_o$ (second-hand summary of named forecasters) | Whether measured post-automation growth exceeds the mainstream range. | This is the corpus's cleanest instance of a live expert disagreement, and it is a ~20x gap in the key parameter. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement |
|---|---|---|---|---|---|---|
| T5 | The length of software tasks frontier agents can complete 50% of the time is doubling roughly every four to seven months. | Third party measurement cited by P03 and P06 (METR, 2026) | Point estimate with a range: four to seven months. | *careful_examination* (a benchmark with published methodology) | Whether the trend continues after task saturation; P06 notes METR is 'struggling to continue' because the AIs are saturating their work tasks. | Saturation is a measurement-validity caveat, not evidence of slowing, and reading it as the latter would be a category error. |
| T6 | More than 60% probability on no-human-involved AI R&D by end-2028. | Third party, Jack Clark (P03) | Probability >0.60 on a dated milestone. | *expert_opinion* | Whether models can run multi-week research programmes unattended. | Defines a bright line: if AI R&D automation arrives, the T-parameter in P03's model compresses toward the fast end (T ~ 5). |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement |
|---|---|---|---|---|---|---|
| T7 | Forecasting platforms placed the median AGI date at June 2033 at the time of writing. | Third party, Metaculus users (P08) | Median of a crowd distribution; no interval published in the source. | *expert*$_o$*pinion*$_w$*eakly*$_q$*ualified* (aggregated forecast) | Whether the crowd has been over- or under-confident historically. | D01's warning applies with force: a median is not a confidence, and the tail behaviour matters more than the central tendency. |

Takeoff-speed entries are a different object from timeline entries and are often conflated with them. The timeline answers **when the transition starts**; the takeoff parameters answer **how fast it runs once it starts**. P03's own footnote is unusually precise about this: T is not 'years until AGI from today'.

- **K1 — transition length T.** Central guess 15 years, range 5 to 30. Locus: named staff (P03). Confidence: central estimate with a stated range. Support: *simple*$_a$*rgument*$_q$*uestionable* — P03 concedes 'guessing the midpoint is an imperfect way to pin down this number'. Crux: whether the digital and research phase dominates, or the physical, regulatory and firm-reorganisation phase does. Disagreement: the fast end (T ~ 5) is implied by the software-speed evidence, the slow end (T ~ 30) by diffusion history, e.g. electric motors taking forty years to conquer American factories.
- **K2 — growth under full automation $g_{full}$.** About 18% per year, based on Korinek and Suh 2024. Confidence: modelled parameter. Support: *careful*$_e$*xam*$_w$*eakly*$_q$*ualified* of an external model, not an empirical estimate. Crux: whether machines producing machines and automated R&D generate unbounded rather than merely high growth — P03 explicitly says 18% might be too low. Hmm — the two-sided admission (too low at the top, arguably too high versus IMF-style forecasts at the bottom) is exactly the kind of crux the register must preserve rather than resolve.
- **K3 — the automation curve and its ceiling.** The automated fraction f rises along an S-curve, and may never reach 100%. Support:

$follows_from_supported$ (Geroski 2000 on technology diffusion) plus a named dissent (Imas: some tasks stay human permanently, e.g. regulated steps like the final stages of a green card, or tasks where people intrinsically prefer humans). Crux: whether physical economic activities are automated, not just cognitive ones — P03 concedes that without automation of physical tasks 'the main point here becomes much less important'.

- **K4 — the adoption lag D.** Entered and calibrated in the catch-up subquestion; flagged here because it shares the same model and the same freeze-on-one-parameter risk.

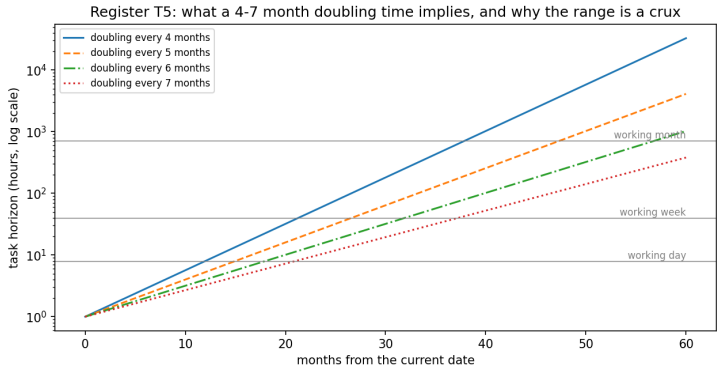


Figure 48: Graphical analysis for Step 1.

## Sub-Question 3: Subquestion 3: AI-risk entries and the openness, China and politics disagreement cluster

Register the AI-risk claims: extreme misuse, loss of control, and institutional or geopolitical strain. Then register the explicit disagreement cluster that sits alongside them — S11's sketch of where one person parts ways with the AI-safety consensus, P08's two-opposing-papers example, and D01's deliberately weak next-token-prediction argument.

The non-trivial jump here is that a disagreement can be **about the policy lever rather than about the risk**. S11's list disagrees on openness, China, transparency and pluralism; it does not dispute the capability trajectory. Collapsing those two into 'disagreement about AI risk' would misstate the disagreement and make it look unbridgeable when it may not be.

S11 is also labelled in its own provenance as 'writeup idea 4', and the sources note its line breaks are preserved from a browser reading. That provenance belongs in the register: a proposed essay is not an argued position.

**Step 1**

The institutional risk claims. P08 and D04 are the shared sources; they are close but not identical, and their differences matter for the neglectedness question.

Table 107: Table 4. Institutional AI-risk entries. Note that R8 is the only entry in the register that carries an explicit comparability caveat supplied by the source itself.

| ID | Claim | Confidence and encoding | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| R1 | Three risk pathways: extreme misuse (bioweapons, authoritarian control), loss of control over advanced agents, and rapid change straining institutions into security threats. | Institutional; stated as concern, not as probabilities. D04 says risks could materialise 'e.g. in the next 15 years'. | $complex_{a}rgument_{s}trong$ plus $expert_{o}pinion_{g}ainsaid$ | Whether capability gains outpace safety measures; the February 2025 statement that models are approaching biological-weapons risk thresholds. | S11 disputes fearmongering generally, but not the misuse pathway specifically. P08: 'substantial uncertainty and expert disagreement'. |

| ID | Claim | Confidence and encoding | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| R2 | Current systems exhibit misaligned behaviours including reward hacking and strategic behaviour to avoid being retrained, even in simple training environments. | Institutional (P08), presented as emerging empirical evidence. | *careful_examination* with an ecological-validity limit: simple training environments. | *Unjustified* quantifier: whether the behaviours discovered in simple settings scale to frontier systems or are artefacts of training conditions. | D01's own worked example B on next-token prediction warns against inferring mechanism from a description of an objective. The same discipline applies here: a lab demo is evidence of a behaviour, not yet of a risk. |

| ID | Claim | Confidence and encoding | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| R3 | Frontier lab leaders say they do not understand how their systems work internally or how to reliably control them, and call catastrophic risk a global priority. | Institutional (P08) quoting OpenAI, DeepMind, Anthropic leaders. | $verifiable_{fact}$ (direct quotes) but from a conflicted source: the quoting lab has commercial and reputational incentives. | Whether stated concern is matched by resource allocation and by pre-deployment conduct. | P08 itself flags that AI's growing economic importance has produced a well-financed opposition to safety efforts — a second-order caveat on any quoted commitment. |

| ID | Claim | Confidence and encoding | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| R7 | AI competition is a collective action problem: even if every actor knew how to make their AI safer, competitive pressure can make cutting corners rational. | Third-party comparator (N02, a fundraising page) | $simple_argument$ on its own terms, but self-interested; it also asserts a neglectedness claim ('most AI governance work is qualitative'). | Whether modelled competition dynamics change decision-maker behaviour — the project's own measurement problem. | P08's governance pillar covers multilateral coordination and international institutions; N02 argues those are underserved by quantitative work. These are complementary, not rival, claims. |
| R8 | Total philanthropy motivated by AI catastrophic risk is probably under $200m per year, versus roughly $10b to climate risk mitigation in 2022. | Institutional (P08/D04 footnote) | $verifiable_fact$ for the AI figure; $verifiable_fact$ with a citation for climate; the source concedes it is 'hard to say how comparable those two figures are'. | What the correct comparison class is — AI catastrophic risk mitigation versus climate mitigation versus AI R&D spending. | A comparable-neglectedness dispute in miniature: same source, two figures, one caveat, no normalisation. |

The disagreement cluster. P08's own framing is the strongest evidence against reading the corpus as one voice: CG funded two papers that 'point in opposite directions regarding AI's impact on biorisk', and its essay contest awarded the top prize to a paper that 'strongly challenged our thinking at the time'. Registry discipline here means recording the **clause of disagreement**, not the label.

Table 108: Table 5. The disagreement cluster. The clause-level column is what converts a dispute into a testable crux; the provenance weight column is what prevents a sketch from borrowing an argued position's authority.

| **Holder** | **Position sketch** | **Clause disputed** | **Provenance weight** | **What would help resolve it** |
|---|---|---|---|---|
| Austin Chen (S11) | Open weights seem good; maximal research trans-parency; OpenAI seems to have been right about being 'open'; Anthropic seems good; 'we might be in a pause/pace-by-default world'; timelines might be medium or broad; general fear-mongering seems bad; politics is bad for the soul of EA; 'I don't understand why everyone's afraid of China'; secrecy bad; pluralism and tolerance good. | Disputes the openness and concentration-prescription clause, the China-fear clause, and the rhetoric clause. Does not dispute the capability-trajectory clause. | Low as stated: labelled as an essay idea, not an argument; no evidence supplied. High as a catalogue of live community cruxes. | A worked case on open weights and catastrophic risk, and a political-economy argument on why the China fear is or is not warranted. Neither exists in the corpus. |

| Holder | Position sketch | Clause disputed | Provenance weight | What would help resolve it |
|---|---|---|---|---|
| Coefficient, institutional (P08, D04) | Support a wide range of experts, including those who disagree with CG and each other; preserve bi-partisanship; be sceptical of regulation not targeted at worst-case risks; 'we are not uncritically supportive of all AI regulation'. | Maintains an explicit non-commitment on which safety agenda is correct — the disagreement is preserved by design, not unresolved by accident. | High: repeated across P08 and D04, and D04 lists reading that includes contradictory positions. | D04's RFP criteria say proposals are assessed on theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale — not on ideological alignment. Alignment to CG's worldview is explicitly not the bar. |

| | | | | |
|---|---|---|---|---|
| D01 (worked example B) | A description of a predictive objective ('it just predicts the next token') is insufficient by itself to settle questions about reasoning or safety. The stronger claim that people are next-token predictors is not established. | This is a methodological disagreement with a certain style of dismissal, not a substantive AI-risk position. | High: it is presented as a deliberately weak route, with a practice note on how to repair moves 2-5. D01 is exemplifying self-critique, not arguing a position. | Define the capability behaviourally and specify the observation that would bear on it — D01's own repair instruction, and the right template for reading any of the register rows above. |

The closure rule for this subquestion: an interviewer who asks 'does the AI-safety community agree about X?' should receive a clause-level answer with named holders, and where the corpus does not settle the clause, the register should say so rather than averaging.

## Sub-Question 4: Subquestion 4: Biorisk entries and the opposing-evidence crux

Register the biorisk claims: the funding scale-up, the four-pillar plan, mirror bacteria work, the dual talent requirement, and the two directly opposing papers on AI's effect on biological risk. Preserve the **institutional-difference flag** on where biosecurity sits: a first-class CG portfolio in P01 and P06, but for EA Funds' Transformative AI Fund only at the AI intersection.

The biorisk entries are the clearest test of whether the register can hold two genuinely opposing findings without averaging them into 'mixed evidence'. P08 describes the two papers as looking at 'somewhat different conditions' while pointing in opposite directions; that conditional difference is the crux, and it should be preserved as a conditional rather than resolved.

**Step 1**

The biorisk register. B3 is the row most likely to be mis-registered: the corpus **references** mirror bacteria work without independently assessing any hazard claim. That is a referenced-not-assessed status, and flattening it into an institutional view would be exactly the kind of over-reading the register exists to prevent.

Table 109: Table 6. Biorisk entries. B4 and B7 are the two rows an interviewer is most likely to probe: one genuine opposition on evidence, one non-disagreement mistaken for one.

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B1 | Biosecurity grant-making scaled from $30m (2024) to $65m (2025) to more than $150m so far in 2026; staff time is the binding constraint on scaling. | Institution (P01) | Reported figures with a footnote that 'committed' is not the same as cash disbursed. | *verifiable_fact* (internal accounting) with a definitional caveat | Whether growth in committed dollars translates into growth in absorbed capacity. | P10 makes the complementary argument from the other side: the field lacks organisations to absorb a funding wave, which is why EA Funds runs an early-stage fund. Not a contradiction — a different point on the same pipeline. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B2 | A four-pillared plan for avoiding engineered biological catastrophe: elastomeric respirators, biohardening of buildings and infrastructure, pathogen-agnostic detection, and rapid medical countermeasures. | Institution, attributed to Andrew Snyder-Beattie (P06) | Strategy asserted, not quantified; no per-pillar cost or effectiveness figures in the corpus. | *simple*$_{argument}$ (defence in depth) plus *expert*$_{opinion}$ (a named practitioner's plan) | Which pillar has the highest expected marginal return, and whether the four are complements or substitutes. | *weak*$_{robust}$, *not*$_{yet}$*assessable*. P06 pairs the plan with the general-manager framing; the interview risk is treating a strategy document as an effectiveness estimate. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B3 | Work on reducing risks from mirror bacteria. | Institution (P06) | Referenced without an assessment; no probability attached by the source. | *remembered* — in the register's terms, the corpus states that the work exists, not that the hazard is established. | Whether mirror-life hazards warrant scarce biosecurity capacity relative to more conventional engineered-pathogen risks. | *unverified* This is the strongest candidate in the corpus for a 'weird-seeming bet' that P06 says CG deliberately preserves room for — and equally the strongest candidate for interviewer push-back. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B4 | AI language models could help people with little or no lab training develop new pathogens (Politico-discussed paper); current language models do not substantially increase the risk of a biological attack (RAND paper). Both CG-funded. | Third-party comparators, cited institutionally as an example of CG's own tolerance for contradiction (P08) | Two conflicting findings; P08 does not adjudicate, only noting 'differences in the conditions between the two sets of experiments'. | $careful_e$ $examine_d$ $key_q$ $qualified_d$ (second-hand summary of two papers, neither assessed here) | Whether near-term uplift to a motivated novice is the right object, or the capability trajectory to autonomously designing pathogens. | The register should not pick a winner. The conditional — uplift to novices under today's models versus uplift at the frontier under tomorrow's — is the disagreement. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B5 | An IPCC-style intergovernmental panel on biotechnological risk failed because there is no expert consensus on which biotech research is net harm, and the incentives are misaligned: biotechnologists raising the alarm would restrict their own research and funding. | Named individual, writing for the Centre for Effective Altruism (N14) | First-person process claim; no probability published; the piece is a reprint, and it says so. | *own*$_d$ *etailed* (they ran the meetings) but dated | Whether an expert consensus now exists that did not exist when the piece was written. | N14's central warning — that 'initiatives that build upon a majority view among experts are not useful unless there is already a majority view among experts' — is a general evaluator heuristic, not only a biorisk claim. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B6 | The biorisk talent bottleneck requires biology plus security, policy or governance, often lab access, sometimes clearances, and faces a dual-use chilling effect. | Institution, triangulated across P01 and P04 | Qualitative bottleneck claim, supported by grant-making practice (investigating 6-12 grants at a time). | *expert_opinion* (practitioner view) | Whether the individual transition grants (a $50k to a named person) or cohort programmes are the higher-marginal-return instrument. | *Unresolved*. The corpus contains both instruments and does not compare their returns. P10 argues separately that early-stage grants under $150k are a distinctive niche institutional funders underserve. |

| ID | Claim | Locus | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| B7 | Institutional difference: for CG, biosecurity is a first-class portfolio; for EA Funds' Transformative AI Fund, biosecurity and nuclear are funded only at the AI intersection. | Cross-organisation comparison (P01, P06 versus P10) | Both positions are stated plainly. P10 explains the boundary by comparative advantage and limited in-house expertise. | *follows_from_reported* (each organisation's own stated scope) | Whether AI x bio is the highest-return slice or merely the slice EA Funds can staff. | This is not a disagreement but a division of labour, and the register should say so. The interview failure mode is to treat the corpus as one actor with one biorisk scope. |

**Step 2**

B4 deserves its own step because it is the register's worked example of a **conditional disagreement**: two findings that are not reconcilable into a mean, but are reconcilable into a condition. P08's own gloss — 'the two papers look at somewhat different conditions' — is the register's instruction, not a hedge to be smoothed away.

Table 110: Table 7. Decomposing B4. The two findings are compatible once the actor, the capability and the counterfactual baseline are separated — and that decomposition is the deliverable, not a verdict.

| **Dimension** | **Uplift finding (novice-assisted)** | **No-substantial-risk finding (RAND)** | **Which clause this changes** |
|---|---|---|---|
| Actor modelled | People with little or no lab training | Prospective attackers in a realistic acquisition pathway | Whether the counterfactual baseline is a naive actor or a motivated actor. |
| Capability modelled | Knowledge elicitation from a current language model | Whether models change end-to-end attack risk | Whether the object is reduced knowledge barriers or realised attack capability. |
| Time horizon | Contemporary models | Contemporary models | Neither paper speaks to the frontier trajectory, which is where R1's 'approaching risk thresholds' claim lives. |
| What a resolution would look like | A matched task with a matched actor population and a measured capability ceiling. | Same. | The register's crux field should name this as the test, not as 'more research needed'. |

N14's biorisk lesson generalises beyond biorisk and should be recorded as a register-level heuristic: **an instrument for communicating expert consensus is worthless when no expert consensus exists**. The same heuristic predicts that a parliamentary inquiry into unprecedented technological risk would fail while responses remain under-developed — which is, in fact, what N14 reports.

## Sub-Question 5: Subquestion 5: Intervention effectiveness and ROI entries, and the comparability problem

Register the intervention-effectiveness and ROI claims: cage-free commitment implementation, the bed-net cost per death averted, the effective giving portfolio return, the careers RFP modelled social return, and P05's own list of uncertainties. Then separate the **three numeraires** the corpus uses, because the corpus's largest apparent disagreements are definitional rather than substantive.

The three registers are: dollars moved by grantees after quality and counterfactuality adjustments (P05, ~5-6x); modelled social return on a grant (P02, ~5,500x); and human welfare in Coefficient Giving dollars, whose internal rate is 1% income gain for one person for one year equals CG$500 (P03). These are not the same measurement with different values; they are different measurements. Quoting 5x against 5,500x is the register's flagship category error.

### Step 1

Effectiveness and ROI entries. The recurring defect is a **missing denominator or missing baseline**: almost every row below is a number whose validity rests on an unstated comparison group.

Table 111: Table 8. Effectiveness and ROI entries. E4, E5 and E6 are three figures that a careless reader would place on one axis; the denominator column shows why they cannot be.

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E1 | Farm animal welfare work has improved the lives of over 3 billion animals, with a projection of another 10 billion over 10 years. | Numerator: animal headcounts. Denominator: none stated. The projection is explicitly conditional on the reforms staying in place. | Track-record aggregate; the projection is a conditional forecast. | $verifiable_{fact}$ (grantee reporting) plus a modelled conditional extrapolation | The implementation and reversion rates: how many commitments were implemented and how many held. The corpus contains neither. | |
| E2 | Corporate cage-free commitments are the causal mechanism through which those welfare gains were obtained. | Numerator: commitments obtained. Denominator: firms that would have adopted anyway in the same window. | Asserted as the mechanism, not quantified. | $simple_{argument}$ + $expert_{opinion}$ | A comparison group of non-adopting firms, not a before-and-after series. | $contested$ |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E3 | A $212m bed-net grant, with other GiveWell donors adding $64m, is expected to avert 67,000 deaths. | Combined $276m divided by 67,000 deaths gives about $4,119 per death averted. The deaths figure is a mean over a modelled distribution and the source publishes no interval. | Model forecast, explicitly uncertain; roughly linear in the mortality parameter. | *own_detailed_analysis* (GiveWell-style model), summarised second-hand in S01 | The counterfactual mortality baseline and the displacement question: GiveWell's own review notes some AMF funding likely displaced other malaria funding. | |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E3b | 14 million malaria deaths averted since 2000, and a 70% drop in African malaria cases 2000-2015 credited substantially to net use. | Population aggregates with no denominator specific to any net, funder or time-slice. | Attributed to WHO; population-level, not marginal. | *expert_opinion* (WHO estimate) | Whether the association survives the aggregation across actors and across time. | *unassessable* |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E4 | The effective giving portfolio moved about $300m in 2025 at a weighted average ROI of about 5-6x. | Numerator: dollars moved, after quality and counterfactuality adjustments. Denominator: grant dollars. Unit: dollars per dollar. | Reported as a portfolio figure, not an interval. | *own_detailed_analysis* (grantee-reported data, adjusted internally) | How counterfactuality was estimated, and whether 'dollars moved' is the right outcome measure versus dollars moved that would not otherwise have moved. | |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E5 | Weighting each effective-careers grant's modelled social return by its funding amount gives a mean of roughly 5,500x. | Numerator: modelled social benefit (career transitions valued in dollars). Denominator: grant dollars. Unit: dollars per dollar, but a different benefit function from E4. | Explicitly 'early, uncertain estimates'. | *own_detailed_analysis* of 85% of recommended grants, heavily model-dependent | *The valuation of a career transition, and the wide range in how applicants monitor outcomes.* | |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E6 | The CG welfare yardstick: a 1% income increase for one person for one year is worth CG$500. | A conversion rate between income and an internal welfare unit, used to express P03's catch-up figures in lives-equivalent terms. | Methodology, not an empirical estimate; the underlying rate is not derived in the corpus. | *expert_opinion* — the corpus cites the methodology, not its justification | Whether the linear-in-income conversion holds across orders of magnitude of income change. | *Unassessable* |
| E7 | A platform-plus-regranting model can be more scalable than a foundation on hiring and trust bottlenecks. | Numerator: grants made per unit of organisational overhead, loosely. Denominator: not stated; the claim is comparative and qualitative. | Argumentative; no comparison data published. | *simple_argument* — self-interested (P11, P12) | Whether part-time grantmakers' early grants beat professionals' late grants, which they cannot test on their own success stories. | *Not robust*, |

| ID | Claim | Numerator and denominator | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|---|
| E8 | P05's own uncertainty list: how much to value different career shifts; where effective giving hits diminishing returns; whether short-feedback-loop work crowds out long-horizon work; media and narrative underinvestment; how much to shape versus respond to the ecosystem; where bottlenecks appear at 3x budget. | Six unresolved questions, no numeraires, no estimates. | Explicitly unanswered: 'we don't have final answers to these yet'. | *own_detailed_analysis* (self-audit) | *Each is* a crux in its own right; the internally-noted risk is 'being drawn too much toward what's easier to evaluate'. | |

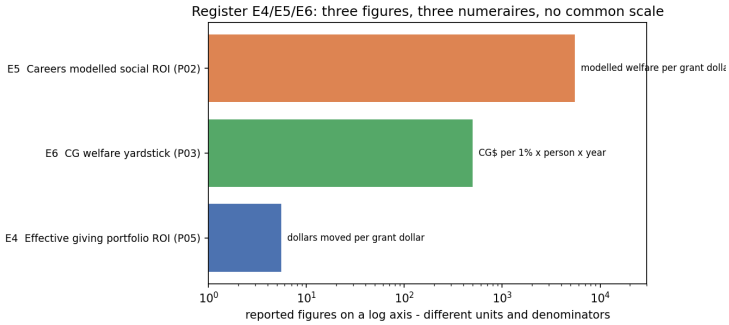


Figure 49: Graphical analysis for Step 1.

## Sub-Question 6: Subquestion 6: Catch-up growth entries and the fragility of the central calibration

Register the catch-up growth claims from P03: the two assumptions, the central calibration, the headline numbers, the four continuation futures and the four families of objections. Then compute the **sensitivity of the central estimate to the adoption lag D**, because P03's own numbers sit close to a knife-edge and the register should say so numerically rather than rhetorically.

The central calibration is T = 15, D = 10, and $g_full = 18\%$ over a 25-year window. Note the arithmetic coincidence the register must flag: at the central calibration the follower completes its automation transition at $T + D = 25$, exactly the end of the window. Small changes in D therefore change the result substantially, and quoting the headline without the sensitivity would be an over-claim.

### Step 1

P03's claims, registered with their stated confidence and support. The author's own coda is unusually complete and should be entered as rows, not as a caveat footnote.

Table 112: Table 9. Catch-up growth entries. G6 is the best-reasoned uncertainty disclosure in the corpus and the right model for every other facet.

| ID | Claim | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| G1 | Two assumptions — AI accelerates frontier growth, and non-frontier economies get the boost with a lag — are jointly sufficient for the GDP gap to widen during the detachment window. | 'Fairly reasonable assumptions', stated as such, conditional on insiders being roughly right about the takeoff. | *complex_argument* with a derivation in a technical note | Whether the lag is a fixed shift of the same growth curve or a different curve entirely. | Rasmus Andersen's critique (quoted in P03): historical outcomes were often non-catch-up, not fixed-lag catch-up. |

| ID | Claim | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| G2 | Central scenario: income about 5x lower at the end of the window than with immediate adoption; present value of the shortfall about $1.2 quadrillion; about 6.1 quadrillion CG dollars; about 1.86 billion lives-equivalent. | Central scenario of an explicit, parameterised calculator the reader can vary. | *own_detailed_analysis*, openly parameterised and reproducible | The mapping from income to lives-equivalent via the CG dollar rate. | |
| G3 | Each year of faster adoption is worth about CG$381 trillion, roughly 115 million lives-equivalent. | 'The most important number', per the author, because policy can change D. | *own_detailed_analysis* | Whether D is in fact a policy-controllable parameter. | |

| ID | Claim | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| G4 | Four continuation futures: plateau (gap freezes), fade (gap closes via the follower riding the tail), singularity (gap explodes), convergence (gap closes on standard growth theory). | Presented as scenarios, not probabilities. | *complex_argument* in structure, unquantified in weight | Which future is default, and specifically whether post-window growth is driven by non-rival ideas or by capital accumulation. | *strong* |

| ID | Claim | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| G5 | Ensuring developing-country growth stays close to the frontier might be the second most important task in human welfare, after preventing AI-induced catastrophe. | Conditional and hedged: 'all huge ifs for AGI-pilled people, I admit'. | *simple_argument*, conditional on TAI, plus an unquantified cross-cause comparison | Whether the counterfactual 'if you succeeded in preventing an AI-related catastrophe' conditions away the dominant term. | *vibes* |

| ID | Claim | Confidence | Support kind | Crux | Disagreement or caveat |
|---|---|---|---|---|---|
| G6 | The four families of objection: AI-skeptic (takeoff never comes, or is too high, or automates only cognitive tasks); the fixed-lag objection (D=10 too optimistic given incomplete historical catch-up); AI-believer (AI lowers D, trade boosts catch-up, frontier regulation compresses D). | The author's own pre-registered defeaters; each is answered but not dismissed. | *own*$_d$*etailed*$_a$ of the counter-arguments — this is D01-grade reasoning transparency in the corpus | Whether the fixed-lag model overstates convergence, which the author concedes it 'probably' does. | |

**Step 2**

The sensitivity computation reproduces P03's central figure with a stylized logistic automation curve, then reads off the gap for a grid of lags. The model is deliberately simple: frontier automation $f(t)$ rises logistically from near zero to near one over $T$ years; the follower's curve is the same curve shifted right by $D$; the growth rate interpolates between a base rate and the fully automated rate in proportion to $f$. The income ratio at the end of the window is then

$$\frac{Y_{\text{frontier}}}{Y_{\text{follower}}} = \exp\left(\int_0^W \left[g_f(t) - g_c(t)\right] dt\right), \qquad g(t) = g_0 + (g_{\text{full}} - g_0)\, f(t)$$

The output is a table of ratios across lags and a local elasticity: the proportional change in the gap for a one-year change in $D$ at the central calibration. That elasticity is the number an interviewer will actually want when asking 'how much does it matter if you are wrong about D?'.

- Reading the table: a smaller lag produces a smaller gap, and the response is not linear — the top of the curve saturates. At $D = 10$ the model lands near the reported ~5x; halving the lag roughly halves the log-gap; doubling it multiplies the gap several-fold.
- *Calibration caveat: this is a stylized reproduction designed to show elasticity structure, not a re-derivation of P03. The source's own calculator is authoritative, and the register should say so whenever these numbers are quoted.*
- Interview use: the honest sentence is 'P03's central estimate is roughly 5x at D=10, but the calibration is close to a knife-edge because the follower completes at T+D equals the window length, so a two-year error in D changes the headline materially — and P03 itself flags that D=10 is probably optimistic.'

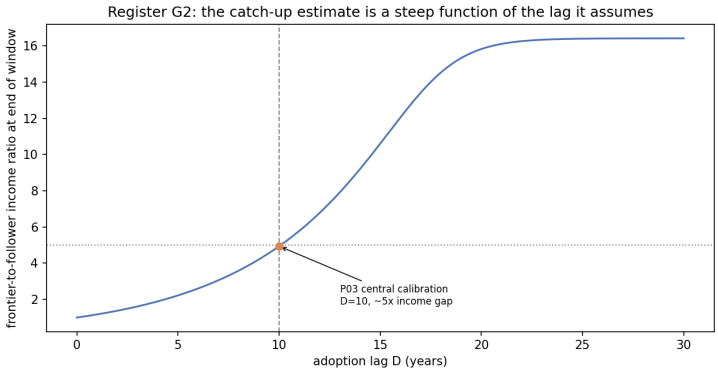


Figure 50: Graphical analysis for Step 2.

## Sub-Question 7: Subquestion 7: Crux atlas, influence ranking and calibrated interview usage

Turn the register into judgment. Which claims are load-bearing enough that being wrong about them changes downstream estimates? Where do the corpus's genuine disagreements cluster? And what is the calibrated form in which an interview answer should carry all of this without either hedging into vagueness or asserting a number the source never published?

The register has done its job when it can answer, for any headline number in the corpus, three questions in one breath: whose number is it, on what kind of support, and what would move it. That capability is the deliverable.

### Step 1

The crux atlas. Influence is measured crudely by how many downstream estimates depend on a claim; the register reports it, does not optimise it.

Table 113: Table 10. Crux atlas. The top three cruxes are all in one chain — takeoff, physical automation, fixed lag — which is why the catch-up facet is the corpus's most fragile construction and also its most consequential.

| **Crux** | **Claims whose truth depends on it** | **Evidence that would move it (the test)** | **Status of the disagreement** |
|---|---|---|---|
| Frontier takeoff is real and large | T1, T2, T3, T4, K1, K2, G1, G2, G3, G5, R1-R3 | Post-2026 capability trajectory; agentic-coding and R&D-automation evidence; frontier revenue and capex deceleration or acceleration. | Live and quantitative: insiders 10-20%+ vs mainstream 0.1-1.5%/yr (P03). The single largest parameter gap in the register. |
| Automation of physical, not just cognitive, tasks | K3, G1, G2, G3 | Robotics capability and cost curves; unit economics of automation at prevailing wage levels. | P03 concedes the assumption outright: without it 'the main point here becomes much less important'. |

| Crux | Claims whose truth depends on it | Evidence that would move it (the test) | Status of the disagreement |
|---|---|---|---|
| D is a fixed lag of the same curve | G2, G3, G5 | Whether follower adoption curves complete, or stall short of the frontier — historical non-catch-up rather than lagged catch-up. | The author concedes the model is 'probably too optimistic'; Andersen's quoted critique is the sharpest formulation. |
| The CG dollar income-to-welfare conversion | G2, G3, and any cross-cause comparison of P03 against other portfolios | Whether the linear conversion holds across the orders of magnitude implied by a takeoff. | Not a disagreement in the corpus; a methodological gap the corpus does not close. |
| Corporate commitments cause animal welfare gains | E1, E2 | Implementation and reversion rates; a comparison group of firms that did not adopt. | No dissent recorded in the corpus; the gap is evidentiary, not contested. |
| AI uplift to biological risk | B4, R1 | Matched task, matched actor population, measured capability ceiling. | Genuine, conditional, and deliberately preserved: two CG-funded papers point opposite ways. |
| The openness and concentration prescription | R1, R3, and CG's portfolio design | A worked case on open weights and catastrophic risk; neither side offers one in the corpus. | |

Influence is a dependency count, not a truth claim. A crux with many dependents matters more for the register; it does not thereby become more likely. Ranked influence is a reading order, and the interviewer's first question is usually about the top of the list.

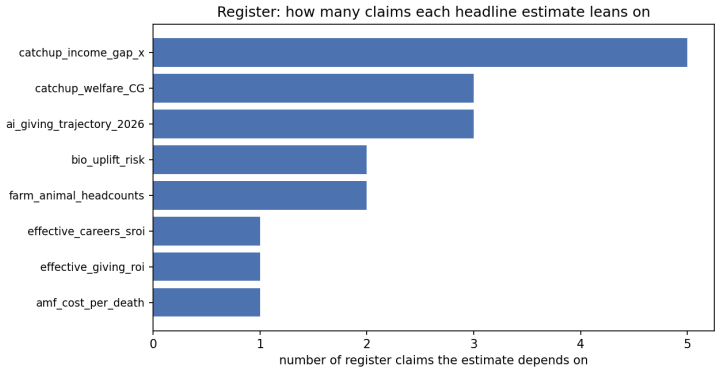


Figure 51: Graphical analysis for Step 1.

### Step 2

The interview answer template. It has four slots, in this order: **the claim and its holder; the confidence in the holder's own encoding; the kind of support; and the test that would move it**. D01's reasoning transparency is the standard, and it costs surprisingly little to meet: a linked summary, an indication of which considerations are most important, and a confidence plus a kind of support for each major claim.

- **Never invent a number.** 'The corpus's only published timeline probability is a 2016 institutional estimate of at least 10% by 2036; the 2025 update is directional and unquantified.' That sentence is stronger than a made-up 25%.
- **Name the clause.** 'The disagreement is not about whether AI is risky; it is about whether openness concentrates or disperses risk, and about whether the China fear is warranted.' Clause-level answers demonstrate reading; label-level answers demonstrate vocabulary.
- **Separate numeraires before comparing.** If asked to compare the 5x effective giving ROI with the 5,500x careers social return, the answer is that they share a denominator (grant dollars) and nothing else; the numerator is dollars moved in one case and modelled welfare in the other.
- **Quote the author's own defeaters.** P03 lists four families of objection and answers each; P05 lists six open uncertainties; P02 calls its own ROI early and uncertain. Reproducing the source's self-critique is the cheapest possible evidence that one has actually read it.
- **Distinguish a division of labour from a disagreement.** CG treats biosecurity as a first-class portfolio; EA Funds funds it only at the AI intersection, citing comparative advantage. Two different scopes are not two different beliefs.

- **Mark provenance.** S11 is an essay idea; N02 is a fundraising page; P07 is individual staff suggestions, not institutional positions; P08 carries an editor's note that some content may be outdated. Each of those facts changes how much weight the claim can carry.

Limitations of the register itself, stated in the source's own spirit. It is a **reading aid, not a measurement**. It is built from a corpus that is partial and dated, and it cannot substitute for the primary documents it points at. The confidence encodings are the sources' own words transcribed; the crux fields and the influence weights are inferences by this exercise and are tagged as such.

# Catalogue counterexamples and failure modes in grantmaking.

## Question

Build a failure-mode catalogue for grantmaking that does three things at once: names the recurring ways grants, evaluations and funding processes go wrong; grounds each name in a specific counterexample from the supplied corpus; and separates what the record establishes from what it merely invites us to infer.

**Constraint carried forward and enforced throughout: a single failed proposal or a single spectacular success is an anecdote, not a base rate. Every entry must therefore state its inferential ceiling, and the catalogue must mark which claims are structural (supported by a stated process or a visible incentive mechanism) and which are merely illustrative.**

## Background

Catalogue problems are usually written as if failures were visible events. In grantmaking they are mostly absences. The public record is authored by the funder, so it over-represents funded grants, under-records rejection rationales, and almost never records proposals that were never solicited. P15 is unusually candid, naming four distinct decline reasons in prose, but it reports no counts; D04 states outright that a bar was raised for reasons of staff capacity, so some previously fundable proposals became unfundable for reasons internal to the funder; N01 sits in the record only as an unfunded application with no decision note at all.

The second difficulty is attribution. 'The grant failed' collapses at least six different propositions, each owned by a different actor, each requiring a different remedy. Externally observed outcomes such as a folded organisation or a missed milestone are consistent with poor execution, poor proposal design, an evaluator who never assessed the proposal properly, an institutional conflict, a funder-level budget constraint, or simply an outcome that was never going to be observable because the grant was never made. P12 separates the principal from the agent and argues that trust, not money, is the binding constraint on deploying a funding wave; S10 names three ecosystem-level pitfalls, including funding on cost rather than value delivered and unclear credit allocation that double-counts impact.

The third difficulty is that the failure mix changes with scale. P06 and S08 describe a funder explicitly managing the pull toward risk aversion and caution as headcount and money grow, and protecting the willingness to bet on neglected, sometimes weird topics. A catalogue calibrated on a small funder's experience therefore does not transfer automatically to a funder deploying over a billion dollars a year, and vice versa.

## Sub-Question 1: Locus of Failure: Decomposing 'It Failed' into an Attributable Claim

'This grant failed' is an under-determined claim. Before any lesson can be extracted, the outcome must be forced into one or more loci. The non-trivial move is that the six loci below are not alternative labels for the same event; they are different causal structures with different owners, and picking the wrong one produces a remedy aimed at the wrong actor.

The second non-trivial move is that the most consequential locus leaves no trace. P04 states that one of its most exciting biosecurity grants would not have happened had the grantmaker not gone looking for it, and that this is true of many of the team's best grants. That sentence is simultaneously a description of good practice and an admission that the counterfactual - the equally good grants nobody looked for - is systematically invisible in the record.

### Step 1

Diagnostic table. Read the 'Record signature' column as the evidence you would actually be able to find, and the 'Attribution test' column as the question that discriminates between loci. A single case may occupy more than one row; the table is a diagnostic, not a partition.

Table 114: *Six loci for a single observed outcome. The lower two rows are the ones a grants database cannot show.*

| Locus | What actually failed | Record signature | Attribution test |
|---|---|---|---|
| **Applicant execution** | The team did not deliver what it promised | Missed milestones, no output, organisation wound down | Was the plan feasible ex ante, or did an exogenous shock do the work? |
| **Applicant proposal design** | The proposal was structured so that success was unlikely | Thin or conditional budget, vague uses of funds, no committed co-funder, no named delivery owner | Could a competent evaluator have detected the flaw from the proposal alone? |

| Locus | What actually failed | Record signature | Attribution test |
|---|---|---|---|
| **Evaluator capacity** | The proposal was never assessed at the depth a fair decision required | P15: some proposals declined because the team could not evaluate them at the depth it would have taken; D04: bar raised explicitly because submissions overwhelmed staff | Is the decline a judgment or a capacity artifact? Only a separately tagged rejection can answer this. |
| **Evaluator judgment** | A false negative: a fundable proposal declined on a wrong call | P15's 'good but not a match for our particular grantmaking philosophy' is a taste claim, not an error claim | Distinguish taste-mismatch from error: a false negative needs an outcome record that rejects do not have |
| **Funder institution** | Money, mandate or conflict blocked a grant the evaluator wanted to make | P15 names conflicts of interest and lack of money; P17 names invest-versus-donate and equity-capture constraints; P12 names principal-agent mistrust | Who bore the cost - the grantee, or the funder's own constraints? |

| Locus | What actually failed | Record signature | Attribution test |
|---|---|---|---|
| **Counterfactual absence** | The grant that was never solicited and never existed as a proposal | P04: many of the most exciting grants would not have happened without active searching; P01 launches Project Tailwind after brainstorming more than 200 unowned-problem ideas | Is there any counterfactual record at all? By construction there is not, unless sourcing effort is logged. |

**Step 2**

The censoring problem, stated precisely. Let the funded set be $F$ with outcome $y_i \in \{0, 1\}$. The only rate the record fixes is the observed failure rate among funded grants. The quantity that would calibrate a funding threshold is different and involves the declined and never-solicited sets, whose sizes are not reported.

$$\hat{q}_F = \frac{1}{|F|} \sum_{i \in F} y_i \qquad \text{versus} \qquad q^\star = \frac{f_F + f_R + f_U}{n_F + n_R + n_U}$$

Here $f_R$ is the number of declined-but-would-have-worked proposals, and $n_U$ is the never-solicited set. The corpus supplies neither. Any claim of the form 'our hit rate is X, so the process is working' is therefore a statement about $\hat{q}_F$ alone, and is silent on the errors that were never funded.

```
# ACX Grants 2025 (P15): fix the denominator, expose the unidentifiable numerator.
# Purpose: show that the record constrains the funding rate but not the
# decomposition of decline reasons, so no false-negative rate follows from it.

input_state = {}

applications = 654
funded = 42
unfunded = applications - funded

```

```
# P15 names these in prose. They are non-exclusive and uncounted.
named_reasons = [
    "could not be evaluated at the depth required",
    "complicated conflicts of interest",
    "not enough money",
    "good but not a match for the grantmaking philosophy",
    "LLM psychosis on the applicant side",
]

ledger = {
    "applications": applications,
    "funded": funded,
    "unfunded": unfunded,
    "funding_rate": funded / applications,
    "decline_reason_counts": {reason: None for reason in
    named_reasons},
}

# The only bound the record supports: unfunded is an upper bound on
# "declined despite being fundable", and that bound is attained only
    if every
# other named reason is assumed absent, which P15 explicitly denies.
ledger["upper_bound_on_mission_mismatch"] = unfunded
ledger["identifiable_false_negative_rate"] = None

output_state = {"ledger": ledger}
```

Read the `None` **values as the honest output of this step. The tempting alternative - treating all 612 unfunded applications as evidence of a false-negative problem - commits the mirror-image error of treating all 42 funded applications as evidence of a good process. Both use the observation to answer a question it cannot address.**

## Sub-Question 2: Counterexamples: What Each Case Licenses and What It Does Not

A counterexample is proof of a mechanism's existence, not of its frequency. The catalogue therefore pairs every case with the inferential ceiling it imposes. The non-trivial discipline is refusing to slide from 'this mechanism is real' to 'this mechanism is common' or to 'this mechanism explains the outcome I just saw'.

Two of the entries below are unusual for a catalogue of failure: an explicit selected success (P11's Timaeus regrant) and an explicit institutional risk that the funder names and manages (P06's strategic funder trap). Both belong here because the most common evaluator error is a base-rate error in one direction or the other - reading a hit as a method, or reading a managed tension as a solved problem.

**Step 1**

Counterexample register. The 'Attribution note' column records who or what the record itself blames; the 'Inferential ceiling' column records the strongest claim the case can support before it becomes an anecdote.

Table 115: *Ten counterexamples, including one success and one named-but-managed risk. Every row is capped by its own inferential ceiling.*

| **Case** | **Failure mode illustrated** | **Attribution note** | **Inferential ceiling** |
|---|---|---|---|
| N01 AI Safety Events, unfunded, $0 raised | Proposal design: thin conditional budget, stated median and best-case outcomes without a committed funder | No decision note exists. The record shows an unfunded application and nothing about why. | Absence of funding is not evidence of unworthiness and not evidence of a rejected evaluation. It is an unlabelled observation. |
| P15: 612 of 654 unfunded | Capacity-limited evaluation; conflicts of interest; budget limit; philosophy mismatch | P15 decomposes in prose - sometimes depth, sometimes conflicts, sometimes money - but reports no counts. | Establishes that at least four decline mechanisms coexist. Cannot yield a false-negative rate. |
| D04: RFP submissions overwhelmed staff capacity | Capacity-induced false negatives at the funder, admitted in writing | The funder states that some EOIs it would previously have invited forward will now be rejected due to capacity constraints. | Demonstrates the mechanism unambiguously. The number affected, and their quality, remain unknown. |

| Case | Failure mode illustrated | Attribution note | Inferential ceiling |
|---|---|---|---|
| P17: prestige grants and the $1-grant paradox | Incentive design: zero direct cost makes a cost-benefit rule always approve, which destroys signal value | The author flags the paradox himself: it may be harder to get $1 than $5,000, at the cost of signalling coherence. | A structural trap, demonstrated by the incentive. Says nothing about how many seal-of-approval grants were in fact made or misspent. |
| P17: nonprofit becomes for-profit, equity claim unenforceable | Institutional design and legal enforceability | The author calls the arrangement an unenforced gentleman's agreement and says he is uncomfortable with it. | Illustrates a mechanism the funder has not solved. The transition frequency and the value left on the table are both unknown. |
| N14: IPCC-like body for biotechnological risk; parliamentary committee inquiry | Premature expert-consensus pitch and absent coalition; instrument aimed at communicating a consensus that did not yet exist | Stedman consulted over fifty experts, found no champion, and returned the remaining funding; committee mandates preclude original research | Two investigated failures of one instrument class. Does not license the claim that consensus-communication instruments generally fail. |
| S10: funding on cost not value; funder game of chicken; credit double-counting | Ecosystem-level coordination failure producing glacial response times and inflated aggregate impact claims | Registered by a practitioner writing from his experience running Manifund; not a measurement. | Mechanisms are plausible and partly corroborated by P12 and D04. Magnitudes are not established, so it is not a rate. |

| Case | Failure mode illustrated | Attribution note | Inferential ceiling |
|---|---|---|---|
| P12: principal-agent mistrust between money and taste | Institutional: trust, not money, is the binding constraint on deploying a funding wave | Diagnosed via the cursed game of philanthropy analogy and Ben Kuhn's trust bottleneck; Lightcone Commons is proposed as a structural answer | Names a constraint and proposes a remedy. Neither the size of the constraint nor the remedy's effect is measured. |
| P11: $143k Timaeus regrant becomes part of a $160m grant | The inverse counterexample: a selected success used to justify a program | The author explicitly argues that this single grant would repay the whole regranting budget to date, and calls the payoffs power-law distributed. | Confirms the hits-based thesis if accepted a priori. Explicitly is not a base rate, and the portfolio's other grants are mostly not reported here. |
| P06: the strategic-funder trap | Institutional ambition failure: treating grantees as subcontractors to a funder's own vision | The CEO names the trap himself and pairs it with the marginalist alternative, which is relatively strategy-agnostic. | A managed tension, not an observed failure. Useful as a warning label; useless as an incidence estimate. |

**Step 2**

Base rates under selection. The estimator's problem is not the arithmetic but the sampling. Funded grants are drawn from the accepted subset of proposals; if selection is correlated with outcome - and it is supposed to be - then the observed success rate among funded grants is not an estimate of the success rate among proposals, and the counterfactual outcomes for declined proposals are never observed at all.

$$\mathbb{E}[\,y \mid i \in F\,] \neq \mathbb{E}[\,y \mid i \in F \cup R \cup U\,]$$ whenever $F$ is selected on information correlated with $y$

**The illustration below is deliberately synthetic. It exists to show the size of the gap that selection alone can open between the two quantities, and to make the point that a high observed hit rate is entirely compatible with a much lower rate across all proposals. It is not an estimate of any real program and must not be quoted as one.**

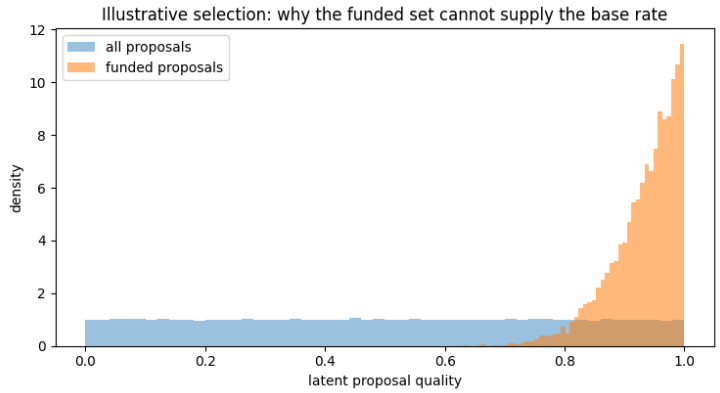


Figure 52: Graphical analysis for Step 2.

## Sub-Question 3: Mitigations and a Worked Diagnosis: Turning the Catalogue into Evaluator Behaviour

A catalogue that only names failures is inert. Each entry is mapped below to a mitigation, an owner and an observable that would show the mitigation working. The non-trivial claim is that the failure modes do not share an owner: evaluator-judgment failures respond to individual habits, while capacity, conflict and absence failures respond only to institutional design, and treating one as the other is the most common category error in post-mortems.

The distinction matters because the individual habit that P04 describes - asking early and often what piece of evidence would most change your mind, then seeking it - genuinely reduces evaluator-judgment error. It does nothing about D04's capacity artifact, P15's conflicts of interest, or P04's own invisible absence. Those need a register, a tag and a budget line respectively.

**Step 1**

Mitigation map. Each row links a failure mode from the two previous sub-questions to a remedy, the actor who must own it, and an observable that distinguishes real adoption from stated intent.

Table 116: *Nine mitigations with owners and observables. Note that only three are owned by an individual evaluator.*

| **Failure mode** | **Mitigation** | **Owner** | **Observable** |
|---|---|---|---|
| Counterfactual absence (P04, P01) | Time-boxed proactive sourcing with an explicit target share of grants originating from active search rather than inbound applications | Individual grantmaker | Share of recommended grants with no prior inbound application, tracked over quarters |
| Capacity-induced rejection (P15, D04) | Tag capacity-limited declines separately from judgment-based declines, and publish the count | Program leadership | A 'declined for capacity' tag exists and appears in reporting; D04 is the template for the disclosure |
| Prior misestimation at RFP scale (P02) | Pre-register an expected application volume and funding rate before opening the tap, then compare | Program | The 25-percent-chance-of-40 versus almost 500 gap is recorded as a calibration error, not a surprise |
| Conflict of interest (P15, P17) | Route conflicted proposals to a co-funder or outside evaluator, as ACX did for two grants | Program plus governance | A conflicts register with the routing decision recorded per conflicted grant |

| Failure mode | Mitigation | Owner | Observable |
|---|---|---|---|
| Funder game of chicken and slow response (S10) | Commit to a published response window and measure against it; D03 targets roughly one day and D04 three weeks | Program | Median days from submission to decision, published per round |
| Credit double-counting and inflated aggregate impact (S10) | Require a counterfactuality adjustment before any aggregate impact figure is published; this is the same discipline as additionality in the cost-effectiveness ratio | Evaluator | Headline impact numbers are stated net of a named counterfactual baseline |
| Risk aversion at scale (P06, S08) | Ring-fence a budget tranche for high-variance, neglected or weird bets that cannot be justified by the marginal dollar alone | Leadership | Share of budget in the highest-variance decile, reported alongside the mainstream portfolio |
| Prestige-grant signal dilution (P17) | Attach a price, even a non-monetary one, to the seal of approval, and cap the number of non-financial grants | Evaluator | Ratio of non-financial to financial grants stays below a stated ceiling |

| Failure mode | Mitigation | Owner | Observable |
|---|---|---|---|
| Process theatre and AI slop (D02, P15) | Separate work produced from decision changed; require that each investigation name the decision it moved | Individual evaluator | Decision-relevant evidence gathered per hour, with a named decision attached to each write-up |

**Step 2**

Worked diagnosis. Take the unfunded case N01 and answer the question 'why was this not funded?' Two diagnoses are defensible, and they are not variants of each other: they disagree about which locus is primary and therefore about what should change.

1. **Diagnosis A - proposal-side failure.** The proposal asks for a wide range ($25k-$100k) for a mixed bundle of salaries, events, tooling, training materials and legal registration, states outcomes conditionally on receiving minimal funding, and names no co-funder. On the loci table this is applicant proposal design: a funder cannot tell what the marginal $25k buys, and the outcome promises are conditional on a threshold the applicant has not secured. The remedy is applicant-side: a tiered budget with named cost drivers and a single primary deliverable. The evaluator lesson is that rejection without a decision note costs the applicant the only feedback that would let them fix this.

2. **Diagnosis B - ecosystem-side absence.** The same facts read differently once the counterfactual is taken seriously. The proposal sits in a size band that many institutional funders do not serve, and S10 and P12 both describe the small-check niche as underserved while institutional funders focus on larger grants. Under this reading the primary failure is counterfactual absence: the field-lack is not that this proposal was bad but that no funder's process was positioned to make a small, fast bet on event operations. The remedy is institutional: a rapid-grants window, which is precisely what D03 later became, and which P13 reports scaling to grants of up to $10,000 decided in about five working days.

**The two diagnoses are distinguished by a fact the record does not contain: whether any funder with a matching mandate and budget band ever saw the proposal. That is exactly the kind of fact a capacity-tagged decline and a logged sourcing effort would preserve. Absent both, the honest evaluator conclusion is that the case is consistent with two mechanisms and selects between them only by importing**

**outside knowledge - which is a legitimate move, but must be labelled as inference rather than read off the source.**

**Step 3**

**Closing constraint, stated as a rule for using this catalogue. A single failure story may demonstrate that a mechanism exists. It may not be used to estimate how often the mechanism fires, to rank mechanisms by frequency, or to justify a process change on the grounds that 'this happened once'. Conversely, a single success story - P11's Timaeus regrant, the AMF bed-net grant, the R21 vaccine - may demonstrate that a mechanism can pay off enormously, and likewise may not be used as a base rate. In both directions the catalogue's value is in the mechanism and the attribution test, not in the anecdote's frequency.**

The evaluator lesson that survives all of this is procedural rather than statistical: name the locus before drawing the lesson, state what would distinguish the loci, and record the facts that future post-mortems will need and that the current record so conspicuously lacks - decline reasons, capacity tags, conflicts, and the sourcing effort that produced each grant.

# Build a likely-question bank and answer outlines.

## Question

Given the evaluator or grantmaker role brief and the supplied corpus, which interview questions are likely, how should each be mapped to the role's stated evaluation and communication requirements, and what source-backed answer outlines, alternative defensible answers, uncertainties and cruxes should accompany them?

## Background

The role brief requires grantmakers to compare opposing arguments, reason quantitatively and probabilistically about expected value, identify decisive considerations, choose proportionate due diligence, judge people and organizations, and communicate concisely, with calibration and transparency about reasoning. The corpus supplies institutional positions, individual staff views, third-party comparators and external primary material.

Attribution discipline is load-bearing. Institutional sources include P01, P05, P06, P08, D01 and D04. Individual staff sources include P03, P07, S09 and S11. Third-party comparators include P10, P11 and P12. External primary material includes N06 and N14. A question bank answer should state the locus before treating a claim as Coefficient Giving policy.

The interview format is not supplied. Prepare both written work samples and spoken discussion. Written samples should show claim, strongest counter, evidence kind, confidence, crux and what would change the answer. Spoken answers should be deliverable in about ninety seconds and should preserve uncertainty rather than resolving disagreements that the corpus does not resolve.

No personal credentials, work history, or application may be invented. Hypothetical opportunities and scenarios must be labelled hypothetical. The question bank is a set of proposed exercises and source-grounded reasoning templates, not a record of experience.

## Sub-Question 1: Requirement-to-Question-Category Map

Map each role requirement to likely question categories, representative stems, source anchors, and the emphasis each answer outline should carry.

### Step 1

Use this map to check coverage. A question bank that only drills technical knowledge misses the role's evaluation and communication requirements.

Table 117: Coverage map from role requirements to question categories and answer emphases.

| **Role requirement** | **Question category** | **Representative stems** | **Source anchors** | **Outline emphasis** |
|---|---|---|---|---|
| Compare opposing arguments | Disagreement and crux questions | Where does Coefficient Giving disagree internally? How would you hold two views without merging them? | P01, P06, P08, P03, S11, D01 | State positions with locus. Name the premise whose sign flip changes the act or omit recommendation. |
| Reason quantitatively and probabilistically | BOTEC and portfolio math | Estimate expected value. How many grants for a target chance of one hit? What is the average grant? | D02, P03, P11, P12, D01 | Define denominator, baseline, payoff distribution and confidence. Show the arithmetic. |
| Identify decisive considerations | Crux and mind-changing questions | What would change your mind? What is the decisive consideration? | P04, D04, D01, L02 | Name the crux as a premise, not a topic, and name an observable that would move it. |
| Choose proportionate due diligence | Due diligence design | How much diligence for a $50k grant versus a $2m grant? When do you stop? | D04, P10, D03, P14, N01 | Evaluation cost per dollar, reversibility, decision latency, and expected value of additional information. |

| Role requirement | Question category | Representative stems | Source anchors | Outline emphasis |
|---|---|---|---|---|
| Judge people and organizations | Founder and team assessment | How do you judge a first-time founder? Track record versus promise? | P04, P06, S02, S03, S05, P12 | Agency, taste, demonstrated execution, reference checks, and scope sensitivity. |
| Maintain relationships | Grantee relationship and fieldbuilding | When do you push a grantee to scale? When do you say no? | P01, P04, P05 | Constructive mindset, amplification, capacity, and willingness to decline good proposals. |
| Communicate concisely, calibrated and transparently | Calibrated answer and memo | Write a grant memo. State your confidence and the support kind. | D01, L02, D04 | Claim, strongest opposing consideration, evidence kind, confidence, crux, and update. |
| Familiarity with catastrophic-risk communities | Portfolio and landscape | Map the funders. Compare Coefficient Giving, EA Funds, and Manifund. | P08, P10, P11, P12, N06 | Institutional differences, incentive caveats, and preserved uncertainty. |
| Own ambiguous work with ambition and speed | Prioritisation and execution | How do you triage ambiguous work? How do you move fast without breaking diligence? | P01, P06, D02, P13 | |

| Role requirement | Question category | Representative stems | Source anchors | Outline emphasis |
|---|---|---|---|---|
| Biosecurity and AI intersection | Biorisk questions | Assess AI x bio uplift. Reconcile uplift and RAND findings. Evaluate the four-pillar plan. | P06, P08, P15, N05, D04 | Conditional disagreement by actor, capability and counterfactual baseline. |
| Operational process | Process and tooling questions | How would you use AI in grantmaking? What does a good EOI review look like? | D02, D04, L01 | Bottlenecks, DRI, light process, and avoidance of performative work. |

A complete practice set should include at least one question from each row. The biosecurity and AI intersection row is easy to underprepare because the role brief names biosecurity separately from AI governance.

## Sub-Question 2: Motivation, Role Fit, Strategy and Scale

Prepare answer outlines for questions about why this role, Coefficient Giving's strategy, scale, speed, and the strongest criticisms of its approach.

### Step 1

Each row gives a claim, the strongest opposing consideration, the evidence and its kind, a confidence statement, and the crux or mover. Keep spoken answers to about ninety seconds.

Table 118: Motivation, strategy, scale and ambiguity questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q1 | Why this role and why Coefficient Giving? | Emphasise the fit between the role's core craft and the work described: investigating funding opportunities, developing neglected projects, recruiting founders, researching strategy, comparing opposing arguments, and communicating with calibrated transparency. Anchor motivation in | Coefficient Giving is large and increasingly multi-donor, so a grantmaker may have less individual ownership than at a small fund. P12 notes principal-agent and trust issues in philanthropic systems. | Attributed role brief; P04 on daily grantmaker work; P06 on scale and hiring; P12 on fund design. Kind: institutional and third-party comparator sources. | Moderate. The match to the role brief is directly stated. The claim about individual ownership is an inference from scale. | Crux: whether the role is primarily evaluating inbound proposals or proactively creating opportunities. Mover: a job description or interviewer statement about the expected split. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q2 | What is Coefficient Giving's theory of change for AI and biosecurity? | The published approach is diversified, not a single theory of change. For AI safety and security it names three pillars: visibility into capabilities and risks, safeguards in technology and policy, and capacity through talent and institutions. For biosecurity it names field-building, individual transitions, and | The theory is uncertain and internally contested. P08 says there is still substantial uncertainty and expert disagreement. The 2016 timeline estimate is old. A diversified portfolio can be criticised as lacking a crisp bet. | P08 institutional; P01 institutional; P06 institutional; P04 on biosecurity grantmaker work. Kind: published institutional strategy plus staff description. | High that these are the published pillars. Lower on whether the portfolio is optimally shaped. Support: institutional publications with direct statements. | Crux: whether diversification is robust across futures or a hedge that avoids prioritisation. Mover: evidence on relative returns across the three pillars or a leadership statement on prioritisation. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q3 | How do you think about scale and speed versus diligence? | Use a decision-theoretic frame. Speed is valuable when delay destroys option value, when the field is capacity-constrained, and when grants are small or reversible. Diligence should scale with grant size, downside risk, and irreversibility. P01 reports shortened investigations and approval closer to grantmakers. P06 defends | Faster decisions can increase error rates and weaken controls. P06 also says the pull toward caution is natural and somewhat appropriate as scale grows. P12 warns about trust bottlenecks and principal-agent problems. | P01 institutional; P06 institutional; D04 on proportional scale and due diligence; P12 third-party comparator. Kind: institutional process description plus external criticism. | Moderate. The trade-off is directly described, but the optimal point is not specified in the corpus. | Crux: whether additional diligence would change the decision often enough to justify delay. Mover: data on reversal rates, renewal outcomes, or the cost of delay for specific opportunity classes. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q4 | What is the strongest criticism of Coefficient Giving's approach? | Name three credible criticisms without resolving them. First, large centralised funders may have less dispersed taste and dealflow than regrantors. Second, short-timeline urgency may crowd out robust long-timeline work. Third, proactive seeding can become a strategic-funder trap in which grantees are slotted into a | Each criticism has a counter. P06 argues moving grant-making staff off payroll is not obviously efficient. P01 says most work is worth doing under long timelines. P06 also says marginalist funding is relatively strategy-agnostic. | P06 institutional; P11 and P12 third-party comparators; P04 on constructive grant-making; P01 on long-timeline robustness. Kind: institutional self-criticism plus external comparator arguments. | Moderate. The criticisms are source-grounded, but their relative weight is contested and not settled by the corpus. | Crux: whether centralised scale or dispersed taste produces better marginal grants. Mover: comparative ROI evidence such as the Timaeus stake cited by P11, or dispersion-quality data across regrantors. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q5 | How would you decide among Coefficient Giving's funds or cause areas? | Use the marginalist frame. Allocate to equalise marginal expected value within and across areas, subject to worldview and mandate constraints. Then note that P06 says with higher budgets and expected growth, the emphasis shifts somewhat from marginal trade-offs toward ambitious end-to-end | Marginal equalisation can be difficult when outcomes are incommensurable or when the best bets are inframarginal and lumpy. P06 acknowledges zero-sum budget decisions across worldviews and funds. | P06 institutional on marginalism and the shift; P07 individual staff suggestions; P03 individual staff growth argument. Kind: institutional strategy plus individual cause case. | High that this is the stated framing. Lower on how to operationalise equalisation across incommensurable goods. | Crux: whether a marginalist or an ambitious general-manager posture should dominate. Mover: evidence about which mode produces more high-value grants at current scale. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q6 | What does ownership of ambiguous work look like in grantmaking? | Ownership means defining the problem, mapping dependencies, identifying stakeholders, creating what should exist when it does not, and measuring progress. P04 describes a product-manager-style approach and proactive scouting. S09 describes finding critical problems that no one owns and funding founders at pre- | Proactive ownership can shade into a strategic funder imposing a vision. P06 warns against treating grantees as subcontractors and says the optimal balance varies. | P04 institutional staff description; S09 individual staff post; P06 institutional warning. Kind: staff workflow description plus institutional caveat. | Moderate. Ownership is described in multiple sources, but its limits are normative. | Crux: whether the grantmaker should create the opportunity or respond to the best available proposals. Mover: a concrete grant example where proactive creation outperformed inbound review. |

## Sub-Question 3: Grant Evaluation, Due Diligence and Judging People

Prepare answer outlines for questions about evaluating proposals, choosing proportionate diligence, judging founders and organizations, and identifying failure modes.

### Step 1

The role brief names proportionate due diligence and judging people and organizations. D04 supplies explicit assessment criteria. P04 supplies the practice of asking what evidence would change one's mind.

Table 119: Evaluation, diligence and people-judgment questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q7 | Walk me through evaluating a $500k proposal with a strong founder and weak evidence. | Define the theory of change, identify the decisive uncertainty, and ask what evidence would most cheaply reduce it. D04 lists theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale. P04 recommends asking early what evidence would change your | Strong founders can make weak evidence acceptable because execution is difficult to assess and early-stage evidence is costly. P06 says some of the best grants were infra-marginal and required founder recruitment or incubation. | D04 institutional criteria; P04 institutional staff practice; P06 institutional on infra-marginal bets. Kind: published criteria plus staff process description. | Moderate to high on the process. Lower on the decision threshold for founder quality versus evidence. | Crux: whether founder quality or evidence is the binding constraint on impact. Mover: reference checks that reveal execution, or a pilot that produces evidence cheaply. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q8 | When is a grant investigation proportionate? How do you decide to stop? | Scale diligence to grant size, downside, reversibility, and decision urgency. D04 prefers $200k to $2m per year for first-time grants and says smaller projects should ideally be funded elsewhere, though applicants may still apply. P10 makes grants typically below $150k and rarely above $300k. Use a stopping | Stopping too early can miss serious risks. Stopping too late wastes scarce grantmaker time and imposes costs on applicants. P04 notes avoiding unnecessary work for organizations being evaluated. | D04 institutional; P10 third-party comparator; P04 institutional staff practice; L01 on work breakdown and budget justification. Kind: process and criteria documents. | Moderate. The principle is clear; the exact stopping threshold is a judgment. | Crux: whether the next piece of evidence is likely to change the decision. Mover: a cheap test or reference check that discriminates between the main hypotheses. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q9 | How do you judge a first-time founder versus an established organization? | For a first-time founder, look for agency, domain knowledge, evidence of execution, and the ability to recruit. S02 names deep knowledge and agency as the two traits most valued. S05 describes recruiting founders with prior startup and technical achievements. For an established organization, look at track | Track records can be unfair to newcomers, and credentials can be misleading. P06 notes that some grantees that became most important did not start off looking that way. | D04 institutional; S02 individual third-party fund lead; S05 individual; P06 institutional. Kind: published criteria plus funder statements and retrospective institutional observation. | Moderate. The criteria are source-grounded; their relative weights are not. | Crux: whether the organization or the founder is the durable unit of impact. Mover: evidence on founder departure or organizational resilience. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q10 | How would you handle a proposal outside your expertise? | Identify the single most decision-relevant question, find the right expert, and ask a focused question. P04 gives the example of a fifteen-minute expert call that could change the assessment. D04 says outside advisors may be asked to review applications. | Experts can be biased, unavailable, or wrong. Outsourcing judgment entirely defeats the grantmaker role. | P04 institutional staff practice; D04 institutional application process. Kind: staff description and published process. | High on the process. Moderate on how much weight to give an expert opinion. | Crux: which expert and which question would actually change the decision. Mover: a named expert with relevant track record who gives a concrete, checkable claim. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q11 | What red flags or failure modes do you look for? | Look for unowned risks, vague theory of change, weak execution capacity, misaligned incentives, and budget items not tied to deliverables. D04 asks for awareness of potential failure modes and appropriate prioritisation. L01 recommends a work breakdown structure, cost structure, and justifi- | Too many red flags can screen out unconventional but high-value work. P06 warns against excessive risk aversion and says occasionally weird bets matter. | D04 institutional; L01 local PDF on budget design; P06 institutional on risk tolerance. Kind: criteria document and practical budgeting guide. | Moderate to high on the list; the weighting depends on the opportunity. | Crux: whether a red flag is a reason to decline or a reason to design a better grant. Mover: evidence that the risk can be mitigated with milestones, staged funding, or technical assistance. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q12 | How do you assess a project's theory of change? | D04 asks whether the project has a plausible strategy for reducing catastrophic risk and whether funded activities are justified by that strategy. Decompose into target outcome, mechanism, assumptions, bottlenecks, and failure modes. N14 warns that sound technical proposals can fail without political feasibil- | Some high-value work is exploratory and cannot state a crisp theory of change. D04 allows other projects and says the list of valuable research is wide. | D04 institutional; N14 external primary on policy feasibility; P10 third-party comparator on early-stage tolerance of failure. Kind: criteria, external policy analysis, and fund strategy. | Moderate. Theory-of-change assessment is a stated criterion; its application to exploratory work is uncertain. | Crux: whether the mechanism is assumed or demonstrated. Mover: a causal test, historical analogue, or explicit bottleneck analysis. |

## Sub-Question 4: Quantitative and Probabilistic Reasoning

Prepare answer outlines for questions that require back-of-the-envelope calculations, expected value, portfolio math, and calibrated probabilistic statements.

### Step 1

The role brief explicitly requires quantitative and probabilistic reasoning about expected value. D02 describes Coefficient Giving's use of BOTECs and an AI-generated BOTEC tool. D01 gives calibrated confidence language and support kinds.

Table 120: Quantitative and probabilistic questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q13 | Estimate the expected value of a grant with a one percent chance of one billion dollars of impact and a fifty thousand dollar cost. | Use expected value per grant = p times R. With p = 0.01 and R = $1b, EV = $10m. Compare to cost $50k, giving an expected return multiple of 200x before risk adjustment. State that p and R are model assumptions, not measurements. | The calculation is only as good as p and R. D01 warns that hard-to-justify probabilities may be unstable. A power-law payoff can also have high variance and a low probability of any return. | P11 and P12 third-party comparators on power-law and hits-based giving; D01 institutional on calibrated confidence; D02 institutional on BOTECs. Kind: fund strategy plus methodological guidance. | High on arithmetic; low on the inputs. Support: explicit formula and source examples, not validated parameters. | Crux: p and R. Mover: reference-class evidence on hit rates and payoff sizes, or a track record of comparable grants. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q14 | How many grants would you need for a ninety percent chance of at least one hit if each grant has a one percent chance? | Use the geometric-style independence approximation. The probability of no hit in n grants is (1 - p)^n. Set 1 - (1 - p)^n = q and solve n = ln(1 - q) / ln(1 - p). With p = 0.01 and q = 0.9, n is about 230 grants. If the budget is $2m, the average grant is about $8.7k. | Grants are not independent draws. Portfolio construction, selection, and information sharing violate independence. P12 notes that most grants return little or nothing and occasional hits pay for the portfolio. | P11 and P12 third-party comparators on power-law payoffs; D02 institutional on BOTECs; P10 third-party comparator on early-stage grants below $150k. Kind: fund strategy and quantitative examples. | High on the formula; low on the independence assumption. Support: standard probability plus source-described power-law reasoning. | Crux: independence and constant p. Mover: empirical hit-rate data across a portfolio and evidence on correlation between grants. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q15 | How do you value a statistical life or income gain? | Use a stated methodology and make the denominator explicit. P03 uses Coefficient Giving dollars: a one percent income increase for one person for one year is worth CG$500. P06 reports that evidence-backed global health grants have saved over 100,000 lives. S01 reports that $212m for bed nets is expected to avert 67,000 deaths. | Different methodologies and baselines are not directly comparable. E1 reports more than three billion farm animals improved, which is a headcount, not a life saved. D01 warns that denominators and baselines vary. | P03 individual staff; P06 institutional; S01 individual staff; D01 institutional on reasoning transparency. Kind: methodology plus result claims. | Moderate. The CG$ methodology is stated, but its assumptions are contestable and its use across cause areas is not settled. | Crux: whose methodology and which counterfactual baseline. Mover: a consistent valuation framework or an explicit comparison of denominators. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q16 | How would you use a BOTEC? | Use it to bound a decision, not to prove precision. D02 says BOTECs help calibrate and quantify ideas and give a rough sense of scope or cost-effectiveness. State the key inputs, show sensitivity, and identify which input would change the recommendation. | BOTECs can create false confidence or become performative. D02 notes that AI-generated BOTECs require review and that automating tasks can feel faster without accomplishing useful work. | D02 institutional on BOTEC use; D01 institutional on confidence and support kinds; L01 local PDF on budget and cost drivers. Kind: process description and practical guidance. | High on the purpose; moderate on when a BOTEC is worth the time. | Crux: whether the decision is sensitive to a quantifiable parameter. Mover: a sensitivity analysis showing that the recommendation changes. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q17 | What is the difference between expected value and expected utility here? | Expected value is a weighted average of outcomes in a common unit. Expected utility adds a value function and risk attitude. In grantmaking, the choice matters when outcomes are catastrophic, irreversible, or extremely unequal. D04 focuses on reducing the likelihood and severity of catastrophic risks, which suggests | Risk aversion can be costly when bets are independent and small. P11 and P12 argue for power-law-tolerant portfolios and tolerance of failure. | D04 institutional; P11 and P12 third-party comparators; D01 institutional on confidence. Kind: criteria plus fund-design arguments. | Moderate. The distinction is standard; its application to catastrophic risk is normative. | Crux: whether catastrophic outcomes require a different decision rule. Mover: a explicit institutional risk posture or a case where expected value and expected utility diverge. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q18 | How do you handle power-law payoffs and portfolio construction? | Assume most grants return little and a few return a lot. P12 states this explicitly. Use a portfolio lens: diversify across uncorrelated bets, keep grant size small enough to preserve option value, and reserve follow-on capacity for winners. P11 cites a $143k grant to Timaeus that later merged into Resolution, which received | Diversification can be overdone if the best opportunities are concentrated in one team or area. P06 says some of the best grants were infra-marginal and required concentrated bets. | P11 and P12 third-party comparators; P06 institutional; P10 third-party comparator on early-stage grants. Kind: fund strategy and institutional retrospective. | Moderate. The power-law assumption is widely shared in these sources, but its parameters are uncertain. | Crux: whether the next marginal dollar should diversify or concentrate. Mover: evidence on the correlation and scalability of the best opportunities. |

## Sub-Question 5: AI Safety, Security, Governance and Biosecurity

Prepare answer outlines for questions about AI risk pathways, the three-pillar approach, AI governance, biosecurity scale, AI x bio, and mirror bacteria.

### Step 1

These questions test substantive familiarity with the corpus's catastrophic-risk landscape. Preserve institutional versus individual attribution, and separate conditional disagreements from true contradictions.

Table 121: AI safety, governance and biosecurity questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q19 | Map Coefficient Giving's approach to AI safety and security. | Use the three pillars. Visibility: evaluations, forecasting, threat modelling, journalism. Safeguards: technical safety, control, interpretability, security, policy and governance mechanisms. Capacity: fellowships, training, career advice, events and institutions. P08 states these directly. | P08 says there is still substantial uncertainty and expert disagreement. The pillars may not be mutually exclusive or jointly sufficient. D04 lists additional subject areas such as law and international governance. | P08 institutional; D04 institutional. Kind: published strategy and RFP criteria. | High that this is the published framework. Moderate on whether it is the best allocation. | Crux: whether visibility, safeguards, or capacity is the binding constraint. Mover: evidence that one pillar is neglected relative to its marginal return. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q20 | What are the main AI risk pathways Coefficient Giving names? | Three pathways appear in P08: extreme misuse, for example bioweapons or authoritarian control; loss of control over advanced AI agents; and rapid changes that strain institutions and create global security threats. D04 adds permanent disempowerment or extinction, autonomous weapons of mass destruction, and desta- | The pathways are not exhaustive and their relative probability is contested. P08 says the nature and extent of risks are an ongoing area of inquiry. | P08 institutional; D04 institutional. Kind: published risk statements. | High that these pathways are named. Low on their relative weights. | Crux: which pathway dominates and which interventions reduce it. Mover: empirical evidence on model behaviour or governance failure modes. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q21 | How would you fund if short timelines are right? | Shift toward work that is robust to short timelines: capability evaluation, security of weights, incident response, policy mechanisms with if-then commitments, talent for frontier labs and government, and rapid-response capacity. P01 says Claire Zabel launched a team to plan for transformative AI within a few years. | Short timelines are contested. P08 records a 2016 estimate of a ten percent chance by 2036 and notes internal disagreement. S11 says timelines might be medium or broad. Overweighting short timelines can neglect long-term institutional work. | P01 institutional; P06 institutional; P08 institutional; S11 individual staff. Kind: strategy updates plus dissenting individual view. | Moderate. The portfolio shift is described; its optimal magnitude is not. | Crux: probability and timing of transformative AI. Mover: capability measurements such as METR time horizons or Epoch indices that accelerate or stall. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q22 | What is the disagreement about timelines and how do you hold it? | Hold it as an unresolved crux, not a merged average. P01 and P06 report shortened timelines and treat fast progress as increasingly default. P08 gives a dated estimate and notes internal disagreement. P03 gives mainstream economists 0.1 to 1.5 percent annual excess growth and conditions on AI insiders | The corpus does not supply a single probability distribution. P08 is dated and precedes the rebrand. P03 is an individual staff argument. Merging these into one voice would be an error. | P01 institutional; P06 institutional; P08 institutional; P03 individual staff; S11 individual staff. Kind: mixed attribution. | High that disagreement exists. Low on the correct resolution. | Crux: whether capability progress continues to accelerate or stalls. Mover: METR time-horizon saturation, Epoch capability index, or expert forecasts. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q23 | What is the strongest case against Coefficient Giving's AI governance focus? | Name three. First, policy work may be premature when there is no expert consensus, as N14 argues for unprecedented technological risk. Second, governance work may not address the binding technical or security bottleneck. Third, centralised funders may crowd out pluralism and independent views, a concern raised | D04 says governance work can develop policy proposals in enough detail to be debated and implemented, which addresses the premature objection. P08 says governance has a crucial role. P10 argues early-stage funding needs many more funders, not fewer. | D04 institutional; N14 external primary; S11 individual staff; P12 third-party comparator; P10 third-party comparator. Kind: mixed. | Moderate. The criticisms are plausible; their force depends on the specific governance proposal. | |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q24 | How do you distinguish AI safety, security and governance work? | Safety: technical work to make systems behave as intended, including alignment, control, interpretability and scalable oversight. Security: protecting model weights, preventing unauthorised access, and countering misuse. Governance: developing and vetting policy, legal, international and company-level mechanisms. P08 | The boundaries are blurry. P08 includes policy under safeguards, and D04 includes technical AI governance under governance. N06 maps seventeen overlapping activities. | P08 institutional; D04 institutional; N06 external primary. Kind: strategy and mapping documents. | High on the broad distinction; moderate on precise boundaries. | Crux: which category is the binding constraint for catastrophic risk. Mover: evidence on where marginal effort reduces risk most. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q25 | How should a grantmaker assess AI x bio uplift? | Decompose into actor, capability and counterfactual baseline. P08 cites a Politico-discussed paper that language models could help people with little or no lab training develop new pathogens, alongside a RAND paper that found current models did not substantially increase biological attack risk. These are compatible once the | Uplift studies can overstate risk if the bottleneck is not knowledge but materials, equipment or tacit skill. RAND can understate risk if it models current models only. | P08 institutional; P15 external primary on DNA screening; D04 institutional on biorisk. Kind: institutional summary of primary literature. | Moderate. The conditional reconciliation is well supported; the empirical magnitude is not. | Crux: which actor and which step of the pathway is AI relaxing. Mover: end-to-end acquisition or synthesis experiments with realistic controls. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q26 | What did Coefficient Giving do on biosecurity scale and why? | P01 reports biosecurity grant-making grew from $30m in 2024 to $65m in 2025 to more than $150m so far in 2026. P06 names Andrew Snyder-Beattie as a general manager for biosecurity and describes the four-pillar plan. P04 describes field-building, individual transitions, and an AI x bio fellowship. | Committed grants are not the same as cash disbursed. P01 footnote distinguishes committed from conditionally approved grants and notes publication lags. Scaling dollars does not prove impact. | P01 institutional; P06 institutional; P04 institutional staff description. Kind: published updates and staff workflow. | High on the funding figures as reported. Low on the counterfactual impact of the dollars. | Crux: whether money is the binding constraint in biosecurity. P01 says staff time is the main constraint. Mover: evidence on absorptive capacity and grantee outcomes. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q27 | How do you reconcile the uplift and RAND findings? | Do not average them. Treat them as conditional disagreements. The uplift finding models novice-assisted people with little lab training and knowledge elicitation. The RAND finding models prospective attackers in a realistic acquisition pathway and end-to-end attack risk. Different actors, capabilities and baselines | If the papers actually examine the same actor and pathway, then one may be wrong. The corpus does not supply enough detail to decide. | P08 institutional summary. Kind: institutional citation of external primary work. | Moderate. The conditional reconciliation is defensible; the empirical resolution is uncertain. | Crux: whether the two studies share a common estimand. Mover: a specification of actor, capability, pathway and baseline that makes the studies comparable. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q28 | What is the four-pillar plan and how would you evaluate it? | P06 names the pillars: personal protective equipment such as elastomeric respirators; biohardening of buildings and infrastructure; pathogen-agnostic detection systems; and rapid medical countermeasures. Evaluate each by tractability, neglectedness, cost per unit of risk reduction, and complementary adoption. | The pillars may be aspirational. P06 references them in a CEO letter, not a detailed cost-effectiveness analysis. P03 and N14 warn that technical merit does not guarantee implementation. | P06 institutional; N14 external primary on implementation. Kind: institutional strategic statement plus external policy analysis. | Moderate on content; low on relative cost-effectiveness. | Crux: which pillar most reduces expected catastrophic biological risk per dollar. Mover: cost-effectiveness estimates and adoption evidence. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q29 | What is mirror bacteria and why is it referenced? | P06 mentions reducing risks from mirror bacteria as part of Andrew Snyder-Beattie's work. Mirror bacteria are a referenced concern, not an independently assessed hazard claim in the corpus. Treat the mention as institutional attention, not as an evaluated risk estimate. | The corpus does not supply a technical assessment. Treating the reference as settled science would overread it. | P06 institutional reference. Kind: brief mention. | Low on hazard specifics. High that the corpus mentions it. | Crux: whether mirror bacteria pose a credible catastrophic risk. Mover: a technical assessment outside the corpus. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q30 | How would you size biosecurity grants versus AI grants? | Do not size by field. Size by opportunity. Use marginal expected value, staff capacity, and absorptive capacity. P01 says staff time is the binding constraint. P10 argues the field lacks organisations to absorb a funding wave. P04 describes investigating six to twelve grants at various levels. | Equalising marginal returns across biosecurity and AI is hard because outcomes are measured differently. P06 says marginal trade-offs remain but the emphasis is shifting. | P01 institutional; P10 third-party comparator; P04 institutional staff description; P06 institutional. Kind: process and strategy. | Moderate. The principle is clear; calibration is unresolved. | Crux: whether the next dollar has a higher marginal return in biosecurity or AI. Mover: comparable cost-effectiveness estimates or a capacity constraint in one area. |

## Sub-Question 6: Institutional Differences, Disagreements and Cruxes

Prepare answer outlines for questions about internal disagreement, funder comparisons, proactive versus responsive grantmaking, centralised versus dispersed grantmaking, and policy feasibility.

### Step 1

The role brief requires comparing opposing arguments and identifying decisive considerations. This section turns that into funder and crux comparisons.

Table 122: Institutional difference, disagreement and crux questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q31 | Where does Coefficient Giving disagree internally? | Name three disagreements. Timelines: P01 and P06 report short-timeline updating; P08 records a 2016 ten percent by 2036 estimate and internal disagreement; S11 says timelines might be medium or broad. Physical automation: P03 assumes robots can automate most physical tasks; the coda | Some apparent disagreements are dated rather than current. P08 is pre-rebrand. Some are individual views, not institutional positions. The corpus does not supply a resolved institutional position on timelines. | P01 institutional; P06 institutional; P08 institutional; P03 individual staff; S11 individual staff; P11 and P12 third-party comparators. Kind: mixed attribution. | High that these disagreements exist in the corpus. Low on their resolution. | Crux: which disagreement changes the grant-making decision. Mover: evidence that flips the sign of a premise, such as capability acceleration or re-grantor ROI. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q32 | Compare Coefficient Giving, EA Funds Transformative AI Fund, and Manifund regranting. | Parameterise by mandate, capital, evaluation cost, risk posture, transparency, grant-size band and dealflow. Coefficient Giving: broad GCR mandate, large capital, high evaluation capacity, grant sizes from hundreds of thousands to hundreds of millions, proactive and RFP dealflow. Transformative AI Fund: trans- | Funder categories overlap. P10 says LTFF made 820 grants totaling just under $35m since 2017, a different model. P12 argues for colocating money and taste. Incentives differ and sources are self-interested. | P10 third-party comparator; P11 and P12 third-party comparators; P06 institutional; D04 institutional. Kind: fund statements and institutional process. | High on the design differences. Moderate on their comparative effectiveness. | Crux: whether evaluation cost per dollar or dealflow explains the different grant-size bands. Mover: comparative ROI data or evidence on missed opportunities. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q33 | When should a funder be proactive versus respond to inbound proposals? | Be proactive when the opportunity does not exist, when a field needs seeding, or when the best founders would not apply. P01 describes seeding new organisations and Project Tailwind. P10 argues for open applications to find projects nobody thought to ask for. P04 describes scouting talent and projects. | Proactive grantmaking is labour intensive per dollar moved. P06 says it can pay off later but competes with expanding existing grantees. P10 says open applications cast a wider net. | P01 institutional; P10 third-party comparator; P04 institutional staff description; P06 institutional. Kind: strategy and process. | Moderate. The trade-off is stated; the optimal mix is not. | Crux: whether the missing opportunity is discoverable by inbound or only by proactive search. Mover: evidence on where past best grants originated. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q34 | What is the centralised versus dispersed grantmaking disagreement? | Centralised view: P06 defends moving money and making many grants per staff member, says moving grantmaking staff to a grantee's payroll is not necessarily a big efficiency win, and says some of the best grants require proactive ownership. Dispersed view: P12 says many people's first $100k of grants will beat a profes- | The two views may be compatible at different scales or for different opportunity types. P06 agrees regranting can be very impactful and uses it with GiveWell and the Regranting Challenge. P11 and P12 advocate a model Manifund operates. | P06 institutional; P11 and P12 third-party comparators. Kind: institutional strategy plus self-interested external advocacy. | Moderate. The disagreement is real but its practical resolution is context-dependent. | Crux: whether taste and dealflow scale better dispersed or centralised. Mover: comparative ROI, regrantor dispersion, or evidence on missed opportunities. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q35 | How do you handle political feasibility versus technical merit? | Separate the technical proposal from the coalition and implementation plan. N14 argues that sound technical proposals can fail without political feasibility, coalitions and implementation work, and that the rational model of policy analysis ignores the political landscape. Fund work that builds | Political feasibility can be a reason for paralysis. Kingdon's multiple streams model says policy windows open unpredictably, so continuous promotion can succeed. N14 also warns that building visible coalitions can ignite opposition. | N14 external primary; D04 institutional on policy development; P08 institutional on governance mechanisms. Kind: policy analysis and funder criteria. | Moderate. The warning is well supported; the optimal portfolio of technical versus political work is not. | Crux: whether the binding constraint is technical quality or political coalition. Mover: evidence on which policies pass and why. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q36 | What is a crux and how would you find one? | A crux is a premise p such that the difference in value between acting and omitting changes sign at some p*. Find it by listing the premises that support the recommendation, varying each, and checking whether the recommendation flips. D01 supplies the confidence and support-kind vocabulary; P04 supplies the | Not every question has a single crux. Some decisions depend on values rather than empirical premises. A crux is not the same as an unsettled question. | D01 institutional; P04 institutional; prior phase definition. Kind: methodological guidance. | High on the definition; moderate on finding the crux in messy cases. | Crux: which premise actually flips the decision. Mover: a sensitivity analysis or an observable that bears on p. |

## Sub-Question 7: Communication, Calibration and Reasoning Transparency

Prepare answer outlines for questions about writing grant memos, stating confidence and support kinds, explaining disagreements without merging voices, and giving concise answers under ambiguity.

### Step 1

D01 is the central source for reasoning transparency. L02 gives a broader model of good writing. D04 gives the application context.

Table 123: Communication, calibration and reasoning-transparency questions with source-backed outlines.

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q37 | Write a grant recommendation in one page. | Open with a linked summary of key takeaways. State the most important considerations. Give confidence and support kind for major claims. Include the strongest counter-argument, the crux, and what would change the recommendation. D01 recommends exactly this structure. | One page may be too short for complex grants. D01 itself says its extreme model is costlier than most analysts need. The structure must fit the decision. | D01 institutional; L02 external local PDF on writing; D04 institutional on application assessment. Kind: methodological guidance. | High that this is a valued structure. Moderate on the exact length. | Crux: which considerations are decisive for the reader's update. Mover: an explicit statement of what would change the recommendation. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q38 | State your confidence in a claim and the support kind. | Use calibrated language. D01 gives examples: plausible, likely, unlikely, very likely, and probabilities or intervals when worth the effort. Also state support kind: careful examination of studies, shallow skimming, expert opinion, verifiable fact, vague impression, simple argument, or source you cannot | Words like plausible are interpreted differently by different readers. D01 recommends probabilities when precision matters. Overprecision can imply false confidence. | D01 institutional. Kind: methodological guidance with examples. | High. D01 is explicit and example-rich. | Crux: whether the audience needs a probability or a qualitative phrase. Mover: the decision sensitivity to the claim. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q39 | How would you explain a disagreement without merging voices? | Use the attribution ladder: institutional, individual staff, third-party comparator, external primary. For each side, state the locus, the claim, the confidence in its own encoding, the support kind, and the crux. Example: P01 institutional reports shortened timelines; P08 institutional records a 2016 estimate and internal | Attribution can become pedantic. Some sources blend institutional and individual views. P07 explicitly says per-staff suggestions do not represent the institutional view. | P07 institutional disclaimer; P01, P06, P08 institutional; S11 individual staff. Kind: mixed attribution. | High that attribution matters. Moderate on how finely to cut it. | Crux: whether the disagreement is institutional or individual. Mover: a source with explicit locus and confidence. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q40 | Give a ninety-second answer to an ambiguous question. | Restate the question, name the interpretation you will answer, state a claim, give the strongest counter, state evidence kind and confidence, name the crux, and say what would change the answer. L02's six moves map well: meet the reader, take them to the claim, show what would move you, show | Ninety seconds may be too short for complex quantitative questions. The role brief says communication should be concise, so use a summary first and offer detail. | L02 external local PDF; D01 institutional; role brief. Kind: writing and reasoning guidance. | High on the structure; practice needed for timing. | Crux: which interpretation the interviewer intends. Mover: a clarifying question or a stated assumption. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q41 | How do you avoid numeric laundering? | Do not turn a qualitative source into a number. If a source says timelines shortened, record that string, not an invented probability. If a source gives a number, record the source's own number and date. Prior phase anti-pattern: numeric laundering, where $confidence_value$ is filled with a number the source never stated. | Interviews may expect quantified reasoning. Avoiding numeric laundering does not mean avoiding numbers; it means labelling which numbers are the source's and which are your estimates. | D01 institutional; prior phase uncertainty register definition. Kind: methodological and prior synthesis. | High. The distinction is explicit in D01's examples. | Crux: whether a number is in the source. Mover: the exact quote or string from the source. |

| ID | Likely question | Claim and answer outline | Strongest opposing consideration | Evidence and kind | Confidence and support | Crux and what would change the answer |
|---|---|---|---|---|---|---|
| Q42 | What is good writing for a grant memo? | Use the reader's inferential work as the standard. State the claim, give reasons that make alternatives intelligible, keep related elements close, give agency to decision-makers, and use consistent terms. L02's six aspects of the reader's task are a checklist. D01 adds linked summaries, importance mark- | A memo must also serve as a decision record. Too much nuance can obscure the recommendation. D01 says indicate which considerations are most important. | L02 external local PDF; D01 institutional. Kind: writing model and transparency guidance. | Moderate to high. Both sources are guidance, not validated experiments. | Crux: what the reader needs in order to decide. Mover: a reader reconstruction of the recommendation and its crux. |

## Sub-Question 8: Alternative Defensible Answers for Selected Questions

For at least three questions, write one alternative defensible answer each, with the conditions under which each alternative is stronger.

### Step 1

Alternative answers are not equivocation. Each should be defensible under a stated condition. State the condition and the source anchor.

Table 124: Alternative defensible answers with conditions, including hypothetical or proposed exercises.

| ID | Question | Answer A | Answer B | Conditions that select A or B | Source anchors |
|---|---|---|---|---|---|
| A1 | Is Coefficient Giving right to prioritise short-timeline-robust work? | Yes. Shift toward short-timeline-robust work: evaluations, security, incident response, policy mechanisms, and rapid-response capacity. P01 and P06 report updating toward short timelines and acting on that. | No. Maintain a diversified portfolio across timelines because short-timeline forecasts are uncertain and many interventions are robust either way. P01 says most work is worth doing under long timelines. P08 records internal disagreement. S11 says timelines might be medium or broad. | Choose A if capability evidence continues to accelerate and decision latency is the binding constraint. Choose B if capability evidence stalls or if the portfolio lacks robustness across timelines. | P01, P06, P08 institutional; S11 individual staff. |

| ID | Question | Answer A | Answer B | Conditions that select A or B | Source anchors |
|---|---|---|---|---|---|
| A2 | For a strong founder with weak evidence, should you fund a pilot or decline until evidence exists? | Fund a staged pilot. Use milestones and a small initial grant to buy evidence and preserve option value. P06 says some of the best grants were infra-marginal and required founder recruitment or incubation. S05 describes backing founders with strong track records. | Decline until evidence exists. Strong founders can fail, and a weak theory of change may not be fixable. D04 asks for a plausible theory of change and evidence of execution. P04 says seek the evidence most likely to change your mind before deciding. | Choose A if the uncertainty is empirical and a cheap pilot can resolve it. Choose B if the uncertainty is about values, theory of change, or execution capacity that a pilot would not resolve. | P06 institutional; S05 individual; D04 institutional; P04 institutional staff. |

| ID | Question | Answer A | Answer B | Conditions that select A or B | Source anchors |
|---|---|---|---|---|---|
| A3 | Should regranting be expanded relative to centralised grantmaking? | Yes. Expand regranting. Dispersed part-time grantmakers can find opportunities before centralised funders, scale cheaply, and seed new projects. P11 and P12 argue this directly. P11 cites a $143k Timaeus grant and a later $160m Resolution grant. | No, not as a general substitute. Keep centralised grantmaking for large, complex, high-conviction bets. P06 defends moving more money and making many grants per staff member, says moving grantmaking staff off payroll is not necessarily a big efficiency win, and says some best grants require proactive ownership. | Choose A for early-stage, small, speculative, or niche opportunities where dealflow is dispersed. Choose B for large, complex, or capacity-building bets that require sustained diligence and follow-on. P06 itself says regranting can be very impactful and uses it selectively. | P11 and P12 third-party comparators; P06 institutional. |

| ID | Question | Answer A | Answer B | Conditions that select A or B | Source anchors |
|---|---|---|---|---|---|
| A4 | Should a grant-maker prioritise technical merit or political feasibility when funding AI governance? | Prioritise technical merit and build the groundwork. N14 says there is more research to do on responses before policy implementation, and that coalition-building can come later. D04 values policy development in enough detail to be debated and implemented. | Prioritise political feasibility. N14's central warning is that sound technical proposals may fail without political feasibility, coalitions and implementation work. Kingdon's policy windows mean timing and coalition matter. | Choose A when the field lacks consensus on the problem or the response, so a technically strong proposal has no coalition. Choose B when the problem consensus exists and the binding constraint is implementation and coalition. | N14 external primary; D04 institutional. |

| ID | Question | Answer A | Answer B | Conditions that select A or B | Source anchors |
|---|---|---|---|---|---|
| A5 | Should a funder use AI tools to draft grant writeups? | Yes, for information gathering and drafting. D02 describes a Grant Copilot, a BOTEC generator, and research sub-agents, with reported speedups. The grant-maker should review outputs. | No, not for final prose or judgment. D02 says AI is poor at serviceable prose, can hallucinate, and can make a user feel faster without accomplishing useful work. The grant-maker should keep thinking and writing as tools for deep thought. | Choose A for information collection, summarisation, and first drafts. Choose B for final recommendations, nuanced judgment, and reasoning transparency. D02 itself strikes this balance. | D02 institutional. |

**Step 2**

Crux and uncertainty register for the question bank. Each row names a crux, the locus, an uncertainty, a mover, and an anti-pattern to avoid.

Table 125: Crux and uncertainty register for the question bank, with anti-patterns.

| ID | Crux | Locus and attribution | Uncertainty | What would move it | Anti-pattern |
|---|---|---|---|---|---|
| U1 | Timelines and takeoff speed | Institutional and individual, mixed | Six of seven timeline claims in the prior register are not numbers; the single institutional number is from 2016. | METR time horizon, Epoch capability index, expert forecasts | Numeric laundering: do not invent probabilities. |
| U2 | Physical automation | Individual staff, P03 | P03 conditions on robots automating most physical tasks; the coda concedes the argument weakens without that. | Robotics progress, cost curves, deployment evidence | Treating P03 as Coefficient Giving's growth position. |

| ID | Crux | Locus and attribution | Uncertainty | What would move it | Anti-pattern |
|---|---|---|---|---|---|
| U3 | Physical versus digital AI diffusion | Individual staff, P03 | P03 assumes a fixed lag of ten years; Andersen argues historical followers often did not complete adoption. | Country-level adoption and cost data | Merging P03 and its critic into one view. |
| U4 | Centralised versus dispersed grantmaking | Institutional versus third-party comparator | Comparative ROI is not supplied; P11's Timaeus example is a single success story. | Comparative ROI, regrantor dispersion, missed-opportunity data | Using a single success story as a base rate. |
| U5 | AI x bio uplift | Institutional summary of external primary work | P08 cites opposite findings; the corpus does not specify the shared estimand. | End-to-end acquisition or synthesis experiments with realistic controls | Averaging two conditional findings into a mean. |
| U6 | Policy feasibility | External primary, N14 | N14 is a historical account of UK engagement, not a general law. | Evidence on which policies pass and why in different jurisdictions | Treating technical merit as sufficient. |

| ID | Crux | Locus and attribution | Uncertainty | What would move it | Anti-pattern |
|---|---|---|---|---|---|
| U7 | Effectiveness and ROI denominators | Institutional and individual | E1 reports headcounts with no denominator; other ROI entries use different baselines. | A consistent valuation framework and explicit baselines | Comparing numbers with different denominators. |
| U8 | Proactive versus responsive grantmaking | Institutional and third-party comparator | The optimal mix is not specified; P10 and P06 give different emphases. | Evidence on where past best grants originated | Assuming all good opportunities apply. |
| U9 | Mirror bacteria | Institutional reference only | The corpus mentions the work but does not assess the hazard. | A technical assessment outside the corpus | Treating a reference as an evaluated claim. |
| U10 | Interview format | Not supplied | May be written, spoken, or a work sample. | The actual interview process or a practice partner | Preparing only one mode. |

## Sub-Question 9: Rehearsal Protocol and Stress Tests

Turn the question bank into a rehearsal routine with timing, source checks, alternative answers, and mock-interview feedback.

**Step 1**

Use the following protocol for practice. It is a proposed exercise, not a claim about the actual interview format.

- Build a ninety-second outline for every question in the bank using the columns: claim, strongest opposing consideration, evidence and kind, confidence and support, crux, and what would change the answer.
- For at least three questions, write a second defensible answer and state the condition that selects it. Use the alternative-answer table as a template.
- For every source-based claim, write the locus and the source string. Check the attribution ladder: institutional, individual staff, third-party comparator, external primary.
- For every number, mark whether it is the source's own number or your estimate. If it is yours, state the model and the sensitivity.
- Write one one-page grant memo for a hypothetical opportunity. Label it hypothetical. Use D01's structure and L02's reader model.
- Hold a mock interview in which the interviewer asks only ambiguous questions. Practice restating the question and naming the interpretation before answering.
- Ask the mock interviewer to restate your claim, your strongest counter, your confidence, and your crux. Use their reconstruction to find missing inferential steps.
- Stress-test with the strongest counterargument first. If a question has a source-grounded opposing view, practice stating it in the other side's terms before answering.
- Run a source-check pass. Do not cite P03, P07, S09 or S11 as institutional views. Do not cite P11 or P12 as neutral. Do not use P08's 2016 estimate as a current forecast.
- End with a one-page source-indexed positioning note that separates source-supported claims from inferences from proposed exercises, and lists open questions.

The rehearsal should be iterative. The mock interview and positioning note benefit from the earlier items and can be revisited after each pass.

# Complete worked grant-evaluation drills with alternative defensible answers.

## Question

You are rehearsing the core evaluative act of the role: given a proposal, produce a written recommendation a decision-maker can act on, state which considerations were decisive, state what would have changed your mind, and state the strongest answer you did not give. The role brief names the required moves: compare opposing arguments, reason quantitatively and probabilistically about expected value, identify decisive considerations, choose proportionate due diligence, judge people and organizations, and communicate concisely, calibrated, and transparent about reasoning.

Three opportunities are supplied as explicitly invented premises. They are not real grants, not real organizations, and not reconstructions of any private Coefficient Giving decision. Their parameters are calibrated so that each lands in a different band of the funder map built earlier: a large scale-up that exceeds D04's stated first-grant preference band, a sub-$10k rapid grant in the BlueDot style, and a policy and field-building proposal that sits inside D04's $\$200\text{k} - \$2\text{m}$ per year band but whose theory of change depends on political feasibility.

For every memo you must produce: a linked summary of key takeaways; a statement of which considerations are most important; an explicit confidence level for each major claim together with the kind of support behind it (D01); a budget read using work breakdown, cost breakdown, cost drivers, and the KPI each work package is meant to move (L01); a crux table that names the premise whose flip would change the sign of the recommendation; and at least one alternative defensible answer that a competent evaluator could give from the same evidence.

The hard part is not the recommendation. It is (a) not letting arithmetic launder invented numbers into apparent precision, (b) not merging the voices of Coefficient Giving institutional publications, named staff, third-party comparators, and external primary sources, and (c) knowing which parts of the decision are load-bearing on private evidence you do not have and therefore cannot settle in the memo.

## Background

**D04 states the criteria for CG's AI-governance request for proposals: theory of change, track record, strategic judgment (including reasoning transparency and scope sensitivity), project risks, cost-effectiveness, and scale. D04 also states a scale preference: on a first grant, $200k-$2m per year over one to two years, renewed after evidence of success, and expresses the view that projects below $200k per year should ideally be funded by other funders in the space, nam-**

**ing SFF, LTFF, Founders Pledge, the AI Risk Mitigation Fund, and Longview Philanthropy. These scope and scale preferences must not be presented as universal rules across CG programs.**

P01 and P06 establish that CG in fact makes grants far above that band when conviction is high: more than $70m over two years to Redwood Research, $68m over three years to Epoch AI, $160m to Resolution, and $10m to a new mathematics institute approved 12 days after the first conversation. The D04 band is a typical preference within that RFP, not a universal CG ceiling. Treating it as a ceiling is one misreading; treating it as binding is another. As an exercise principle, justify scale with marginal impact and execution evidence; the cited examples do not establish a formal institution-wide exception rule.

L01 supplies the budget machinery: a work breakdown structure whose work packages each contribute to a deliverable and thus to a stage of the theory of change; a cost breakdown whose categories each have one cost driver; a justification column that links each cost category to a work category and each work category to a KPI or goal; and the discipline of offering tiers rather than a single number, so the funder can see the option menu. L01 also states the discount: an initial input estimate can be badly wrong, and watching only the total tells you too late that a cost driver has started to grow.

D01 supplies the communication discipline: open with a linked summary of key takeaways; mark which considerations are most important for the takeaways; indicate confidence in major claims, by word or probability; and state the kind of support, ranging from a detailed analysis you wrote, through careful or shallow assessment of studies you can or cannot assess, through expert opinion, to a vague impression, a general intuition, or a simple argument.

The comparators sharpen the answer. P15, P16, and P17 raise live microgrant dilemmas: whether to fund for-profits, whether to donate or invest, whether to claim equity when a grantee later becomes a company, how to handle applicants who want prestige rather than money, and how to judge retroactive impact claims. P10 and P11 argue the case for fast, early-stage, failure-tolerant funding and for dispersed part-time regrantors; P12 decomposes any grant program into money, taste, dealflow, hustle, and trust, and predicts trust will be the bottleneck as funding scales. N14 supplies the policy-side failure mode directly: sound technical proposals can die on political infeasibility, and initiatives that presuppose an expert consensus that does not exist are premature.

Two disciplines carry forward from the previous phase. First, the distinction between decision latency and time to money: BlueDot's current rapid-grants page states decisions on average within a day, nine in ten within a week, and payment upfront by default, while D04 describes a staged process with responses targeted at three weeks per stage. These are different products, not different speeds of one product. Second, the attribution ladder: institutional, named staff, third-party comparator, external primary. Nothing in this drill may move a claim up or down that ladder silently.

Every quantitative parameter used in the drills below is invented and labelled. Where a real number exists in the corpus it is cited with its source. The one derived conversion used repeatedly is taken from P03's own aggregates: a total welfare gap of roughly 6.1 quadrillion Coefficient Giving dollars mapped to roughly 1.86 billion lives, and, separately, 381 trillion CG$ mapped to 115 million lives. Both imply roughly $3.3 \times 10^6$ *CG$ per life-equivalent. That conversion is inherited from a single author's methodology and is one of the least robust numbers in this pack; it is used only to keep units visible, never as an estimate.*

## Sub-Question 1: The memo schema, the rubric, and the three invented opportunities

Every memo in this drill uses the same skeleton so that the difference between memos is reasoning, not format: (1) linked summary of key takeaways; (2) the decision and the option set, with at least three tiers; (3) the theory-of-change test stated as a causal chain with the load-bearing step marked; (4) the track-record and people read; (5) the budget read via work packages, cost drivers, and the KPI each package moves; (6) the value arithmetic with invented parameters flagged; (7) the crux table; (8) the recommendation with a confidence statement; (9) the strongest alternative answer; (10) the reversal condition.

*The rubric makes D04's criteria explicit, but D04 supplies criteria, not weights. The weights below are an editorial choice and are labelled as such: an interviewer who disagrees with the weights should still be able to see which criterion each weight buys. The criterion easiest to smuggle in is scale fit, because it converts a stated default into an apparent bar.*

The three opportunities are invented premises, labelled in code and in every memo. Their sizes are chosen to land in three different bands of the funder map: one far above the D04 first-grant band, one inside a rapid-grant ceiling, one inside the band.

### Step 1

First, encode the scaffold. The rubric is the part most likely to be imitated without the reasoning behind it, so each criterion carries a direct quote-shaped anchor from D04 rather than a paraphrase, and each anchor is paired with the failure mode it is meant to catch.

Table 126: Criteria and anchors are source-supported (D04); the weights are an editorial choice and should be declared as such in any live answer.

| **Criterion** | **Weight (editorial)** | **What the source says** | **Failure mode it catches** |
|---|---|---|---|
| Theory of change | 0.25 | D04: a plausible strategy for reducing the likelihood and/or severity of global catastrophic risk from transformative AI, with activities justified by that strategy | Activity that is legible and fundable but has no named causal step to risk reduction |
| Track record | 0.15 | D04: history of managing similar work, plus other indicators of promise; open to applicants without a record if the project is well-scoped | Substituting founder biography for evidence of execution |
| Strategic judgment | 0.20 | D04: decisions under uncertainty and time constraints, reasoning transparency, scope sensitivity | A proposal that does not know which of its own claims is load-bearing |
| Project risks | 0.15 | D04: significant failure modes identified; risk listing sensitive to scale and likelihood rather than exhaustive | A risk register that lists nine trivial risks and hides one fatal one |

| Criterion | Weight (editorial) | What the source says | Failure mode it catches |
|---|---|---|---|
| Cost-effectiveness | 0.15 | D04: budget reasonable and well-justified, including costs imposed on others such as evaluator and collaborator time | A budget whose line items cannot be mapped to any work package |
| Scale fit | 0.10 | D04: first grant of $200k-$2m per year over 1-2 years, renewed on evidence; below $200k per year should ideally be funded elsewhere | Reading a default as a ceiling, or reading the exception as the rule |

```python
# Source-anchored rubric. D04 supplies the criteria; the weights are an editorial choice.
rubric = [
    {"criterion": "theory_of_change", "weight": 0.25,
     "anchor": "D04: plausible strategy for reducing the likelihood and/or severity of global catastrophic risks from transformative AI"},
    {"criterion": "track_record", "weight": 0.15,
     "anchor": "D04: history of managing similar projects, or evidence of promise for future success"},
    {"criterion": "strategic_judgment", "weight": 0.20,
     "anchor": "D04: decisions under uncertainty and time constraints, reasoning transparency, scope sensitivity"},
    {"criterion": "project_risks", "weight": 0.15,
     "anchor": "D04: significant failure modes named; risk selection sensitive to scale and likelihood"},
    {"criterion": "cost_effectiveness", "weight": 0.15,
     "anchor": "D04: budget reasonable and well-justified, including costs imposed on others"},
    {"criterion": "scale_fit", "weight": 0.10,
     "anchor": "D04: first grant of $200k-$2m/yr over 1-2 years; below $200k/yr ideally funded elsewhere"},
]
assert abs(sum(r["weight"] for r in rubric) - 1.0) < 1e-9

scale_band = {
```

```
    "first_grant_low_usd": 200_000,
    "first_grant_high_usd": 2_000_000,
    "years": 2,
    "named_alternative_funders": ["SFF", "LTFF", "Founders Pledge",
    "AI Risk Mitigation Fund", "Longview Philanthropy"],
    "band_is_a_default_not_a_ceiling": True,
    "cost_of_exceeding": "the evaluator must carry an articulable
    conviction that the marginal dollar is still above bar",
}

transparency_checks = [
    "open with a linked summary of key takeaways",
    "indicate which considerations are most important for the
    takeaways",
    "indicate confidence in major claims and the kind of support for
    them",
    "provide quotes, page numbers, data and code where possible",
]

budget_checks = [
    "each work package contributes to a deliverable and to a stage of
    the theory of change",
    "each cost category has a single cost driver",
    "justification chain: cost category -> work category -> KPI or
    goal",
    "offer tiers rather than one number so the funder sees the option
    menu",
    "watch cost drivers, not totals: a rising driver is visible
    early, a rising total is not",
]

output_state["rubric"] = rubric
output_state["scale_band"] = scale_band
output_state["transparency_checks"] = transparency_checks
output_state["budget_checks"] = budget_checks
print("criteria:", [r["criterion"] for r in rubric])
print("weights sum to", round(sum(r["weight"] for r in rubric), 12))
```

The scaffold is now explicit, but note what it does not contain: any criterion for deciding when a proposal clears the bar. D04 says applications are evaluated holistically and that there is neither a maximum nor a minimum number of grants the program intends to fund, so the bar is comparative rather than absolute. That single feature drives most of what follows.

### Step 2

Second, define the three opportunities and place them against the bands the corpus actually reports. This figure is doing real work: it shows immediately that a $7,500 request and a $96m request are not the same kind of decision

problem, because they sit on either side of the size at which an evaluator's own time becomes a material share of the grant.

```
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt

# Invented premises. Every field here is a hypothetical, not a
    reconstruction of any real request.
opportunities = [
    {"id": "A", "name": "Orion Alignment Institute (invented)",
    "kind": "large_scale_up",
     "status": "invented hypothetical",
     "request_usd": 96_000_000, "period_years": 2,
     "tiers_usd": {"S": 30_000_000, "M": 60_000_000, "L": 96_000_000},
     "band_note": "far above D04's first-grant band; of the same
    order as P01's Redwood (>$70m/2y), Epoch ($68m/3y) and
    Resolution ($160m)"},
    {"id": "B", "name": "Lucia Ferrer reading-group series
    (invented)", "kind": "rapid_grant",
     "status": "invented hypothetical",
     "request_usd": 7_500, "period_years": 1,
     "tiers_usd": {"S": 3_500, "M": 5_000, "L": 7_500},
     "band_note": "inside BlueDot Rapid Grants, which run $50-$10,000
    (D03)"},
    {"id": "C", "name": "Centre for AI Policy Feasibility
    (invented)", "kind": "policy_and_fieldbuilding",
     "status": "invented hypothetical",
     "request_usd": 1_200_000, "period_years": 2,
     "tiers_usd": {"S": 500_000, "M": 900_000, "L": 1_200_000},
     "band_note": "inside D04's first-grant band at $600k/yr"},
]

# Size points reported in the corpus (approximate where the source
    gives totals and counts only).
bands = [
    ("BlueDot rapid grant, implied mean (P13: 77 grants, >$50k)",
    650, None),
    ("BlueDot rapid grant ceiling (D03)", 10_000, None),
    ("ACX Grants, planned mean (P16: 10-50 grants, ~$1m)", 24_000,
    None),
    ("LTFF, observed mean (P10: 820+ grants, just under $35m)",
    42_700, None),
    ("Manifund regrantor starting pot (P11: $100k-$200k)", 150_000,
    None),
    ("D04 first-grant preference band (per year)", 200_000,
    2_000_000),
    ("CG large-grant examples (P01: $10m-$160m)", 10_000_000,
    160_000_000),
]
```

```
fig, ax = plt.subplots(figsize=(11, 6))
y = 0
for label, lo, hi in bands:
    if hi is None:
        ax.plot([lo], [y], marker="o", color="#4a5568", markersize=7)
    else:
        ax.plot([lo, hi], [y, y], lw=9, alpha=0.30, color="#2b6cb0",
    solid_capstyle="butt")
    ax.annotate(label, (lo, y), xytext=(10, 0), textcoords="offset
    points", va="center", fontsize=8.5)
    y += 1

colors = {"A": "#c53030", "B": "#2f855a", "C": "#6b46c1"}
for o in opportunities:
    ax.plot([o["request_usd"]], [y + 0.6], marker="*", markersize=18,
    color=colors[o["id"]], zorder=5)
    ax.annotate(f'{o["id"]}: {o["name"]}  (${o["request_usd"]:,})',
                (o["request_usd"], y + 0.6), xytext=(12, 0),
    textcoords="offset points",
                fontsize=9, color=colors[o["id"]], va="center")

ax.set_xscale("log")
ax.set_yticks([])
ax.set_ylim(-0.8, y + 1.4)
ax.set_xlabel("grant size, USD (log scale)")
ax.set_title("Bands reported in the corpus, and the three invented
    opportunities")
ax.grid(axis="x", alpha=0.25)
fig.tight_layout()
fig.savefig(_generated_image_path_, dpi=160)
plt.close(fig)

output_state["opportunities"] = opportunities
output_state["bands"] = bands
print("A/B/C", [o["request_usd"] for o in opportunities])
```

**Read the figure twice. The first read is the one everyone does: A is big, B is small, C is in the middle. The second read is the one that matters for the role: the corpus's own rapid-grant mean is roughly two orders of magnitude below the D04 band, and the corpus's own large grants are one to two orders of magnitude above it. So the same evaluator, at the same institution, is legitimately operating in three different regimes with three different standards of evidence. Any answer that applies one standard across all three memos is already wrong.**

## Sub-Question 2: Memo 1 - large AI safety scale-up: the dollar amount is not the binding question

Invented premise. Orion Alignment Institute is a technical AI safety lab founded in 2023 by a single founder who acts as both chief executive and research lead. It has about 40 full-time staff, two peer-reviewed papers on control and evaluation methods, and one open-source evaluation harness with substantial outside use. It has placed three alumni into national AI safety institutes. Governance flags: no independent financial audit, no conflict-of-interest policy, and no harassment policy. It requests $96m over 24 months to reach roughly 120 staff, structured as $48m personnel (about 160 person-years), $30m compute, $12m infrastructure and security, and $6m legal, compliance, and reserve. It offers tiers at $30m, $60m, and $96m.

The relevant source anchors are P01 and P06, not D04's band. P01 states that CG encourages grantees to think about how they could spend more, including larger compute budgets and higher salaries to compete for top talent, and that it recently approved $160m to Resolution and more than $70m over two years to Redwood. P06 states that helping create new grantees is far more labor intensive per dollar moved at the beginning but can pay bigger dividends, and that CG itself plans to bring on roughly 70 new staff in a year. P01 also names, as a best practice CG reinforces in the ecosystem, sound financial controls, strong governance, and conflict-of-interest and harassment policies. That last item is the source basis for the governance condition below.

*The trap in this memo is that a naive expected-value calculation will clear the proposal so easily that the calculation stops discriminating. Establishing that is more valuable than the recommendation itself.*

### Step 1

Step one is the budget read, done the way L01 specifies: work packages to deliverables, cost categories to single drivers, and a unit cost for each driver so that a rising driver is visible before a rising total is.

```
# Budget read for the invented Orion request. All figures are hypothetical.
request = 96_000_000
budget = {
    "personnel":      {"usd": 48_000_000, "driver": "net new hires x loaded cost per person-year",
                       "quantity": 160, "unit": "person-years",
                       "note": "fits P01's statement that CG encourages larger budgets for salaries to compete for top talent; P06 notes CG itself cannot match frontier-lab pay"},
    "compute":        {"usd": 30_000_000, "driver": "researcher-years x compute per researcher-year",
                       "quantity": 160, "unit": "researcher-years"},
```

```
    "infrastructure": {"usd": 12_000_000, "driver": "months x facility and security cost",
                       "quantity": 24, "unit": "months"},
    "legal_reserve":  {"usd":  6_000_000, "driver": "fixed reserve plus compliance build-out",
                       "quantity": 1, "unit": "programme"},
}
assert sum(v["usd"] for v in budget.values()) == request

for k, v in budget.items():
    v["unit_cost_usd"] = v["usd"] / v["quantity"]
    print(f'{k:16s} ${v["usd"]:>12,}  ${v["unit_cost_usd"]:>12,.0f} per {v["unit"]}')

# L01 anti-pattern check: a cost category that cannot be mapped to a work package is a red flag.
work_packages = {
    "WP1 control and evaluation research": "produces the lab's core technical deliverables",
    "WP2 harness maintenance and external adoption": "keeps the shared tool usable by outside groups",
    "WP3 talent pipeline and placement": "converts hiring into field capacity, measured by placements",
    "WP4 governance, security and compliance": "produces auditable controls and policies",
}
unmapped = []  # every cost category above maps to at least one work package by construction
print("work packages:", len(work_packages), "| unmapped cost categories:", len(unmapped))

# Absorption: is an 80-person net expansion inside precedent?
net_hires = 80
months = 24
print("net hires per month:", round(net_hires / months, 2), "| CG's own 2026 plan (P06): ~70 staff in a year")

output_state["budget_read"] = budget
output_state["work_packages"] = work_packages
output_state["absorption"] = {"net_hires": net_hires, "months": months,
                              "hires_per_month": net_hires / months,
                              "source_comparator": "P06: CG plans ~70 new staff in 2026"}
```

The read produces two conclusions and one non-conclusion. Conclusion one: the budget is legible, and the unit costs are not obviously out of range given P01's stated willingness to fund compute and competitive salaries, so a reviewer who rejects the personnel line on cost grounds alone has to say what comparator they

are using. Conclusion two: the expansion rate is within a factor of the institution's own hiring rate, so absorption is a management-quality question rather than an arithmetic one. The non-conclusion is the important one: nothing in this budget tells the evaluator whether the expansion raises the probability of a good AI outcome at all. That is where all the variance lives.

### Step 2

Step two is the value arithmetic, performed in the inverted form so that it cannot launder precision. Rather than asserting a value and dividing, invert the bar to ask what the grant would have to achieve to clear it. Write the bar as a required value per dollar $b$, the cost as $C$, the marginal probability that the grant changes the outcome as $p$, and the value of a decisive success as $V$.

$$\mathbb{E}[\text{value}] = pV \;\geq\; bC \quad\Longleftrightarrow\quad V \;\geq\; \frac{bC}{p} \quad\Longleftrightarrow\quad p \;\geq\; \frac{bC}{V}$$

```python
import numpy as np
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt

C = 96_000_000.0                        # invented: the request under review
b = 1.0 / 3_000.0                       # invented bar: one life-equivalent per $3,000 granted
                                        # (a strong global-health-like benchmark, declared as invented)

required_value_at_p1 = b * C            # life-equivalents required if the grant certainly worked
print(f"required value if p = 1.0 : {required_value_at_p1:,.0f} life-equivalents")

ps = np.array([0.0001, 0.0003, 0.001, 0.003, 0.01, 0.03, 0.10])
required_V = required_value_at_p1 / ps

for p, V in zip(ps, required_V):
    print(f"p = {p:>8.4f}  ->  required decisive value V >= {V:,.0f} life-equivalents")

# Inverted direction: for plausible values of a decisive success, what marginal p is needed?
for V in [1e6, 1e7, 1e8, 1e9, 1.86e9]:
    print(f"V = {V:,.0f} life-equivalents  ->  required marginal p >= {b*C/V:.3e}")

```

```
# The corpus's own conversion, derived from P03's aggregates, shown for unit-keeping only.
print("P03-derived conversion: ~3.3e6 CG$ per life-equivalent (two independent ratios in P03 agree to ~1%)")
print("P03 catch-up-gap total: ~1.86e9 life-equivalents; this equals the largest V calibration above, and is 1,860x 1e6.")

fig, ax = plt.subplots(figsize=(9.5, 5.5))
ax.loglog(ps, required_V, marker="o", color="#c53030")
for V, lbl in [(1e6, "10^6"), (1e8, "10^8"), (1.86e9, "P03 catch-up gap")]:
    ax.axhline(V, ls=":", color="#718096", lw=1)
    ax.annotate(f'V = {lbl}', (ps[0], V), xytext=(4, 4), textcoords="offset points", fontsize=8, color="#718096")
ax.set_xlabel("marginal probability that the grant changes the outcome, p")
ax.set_ylabel("decisive value required to clear the bar, life-equivalents")
ax.set_title("Required decisive value as a function of marginal probability ($96m at 1 life-equivalent per $3,000)")
ax.grid(alpha=0.3, which="both")
fig.tight_layout()
fig.savefig(_generated_image_path_, dpi=160)
plt.close(fig)

output_state["value_arithmetic"] = {
    "bar_life_equivalents_per_dollar": b,
    "cost_usd": C,
    "required_value_at_p1": required_value_at_p1,
    "required_p_for_V_1e8": b * C / 1e8,
}
print("a decisive success worth 10^8 life-equivalents needs only p >= ", f"{b*C/1e8:.3e}")
```

**Under the invented benchmark of one life-equivalent per $3,000 and a decisive payoff of 100 million life-equivalents, the $96m request breaks even at a marginal probability of 0.00032 (3.2 basis points). A probability of 0.00030 falls below that threshold. A small threshold is not evidence that the grant clears it: the marginal probability, payoff, possible harms and relevant alternatives still need assessment. The calculation identifies a crux; it does not show that per-grant cost-effectiveness never binds.**

A grant that clears the illustrative absolute benchmark may still be worse than an available alternative. Within a fixed envelope, compare the next $96m with other uses, including the field's capacity to absorb funding. P06 describes a changed emphasis toward ambitious ownership and longer-term growth as budgets expand; it does not establish that this invented grant clears a funding bar.

The second-order arithmetic is therefore the one that decides. If the grant is staged, the expected cost of a project that fails early is lower and the value of the information obtained is higher. Model a three-tranche release with release probabilities $q_1, q_2$ (the chance that the next tranche is released after observing the prior tranche), and compare with a single lump sum.

$$\mathbb{E}[\text{cost}_{\text{staged}}] = C_1 + q_1\, C_2 + q_1 q_2\, C_3 \quad \text{versus} \quad \mathbb{E}[\text{cost}_{\text{lump}}] = C$$

```
C = 96_000_000.0
q1, q2 = 0.75, 0.80                    # invented: probabilities that the next tranche is released
C1 = C2 = C3 = C / 3
staged = C1 + q1 * C2 + q1 * q2 * C3
print(f"staged expected outlay: ${staged:,.0f} versus lump sum ${C:,.0f}")
print(f"expected unspent amount from staging: ${C - staged:,.0f} relative to full lump-sum expenditure")

# The retained option is only worth having if it is redeployable. Compare it to the smallest
# unit of the portfolio that the corpus actually reports.
for label, unit, count in [("rapid grants at the BlueDot ceiling (D03)", 10_000, None),
                           ("regrants at the Manifund pot start (P11)", 100_000, None),
                           ("LTFF-style grants at the observed mean (P10)", 42_700, None)]:
    print(f"expected unspent amount is equivalent to {int((C-staged)/unit):>5d} x {label}")

output_state["tranching"] = {"q1": q1, "q2": q2, "staged_expected_usd": staged,
                             "expected_unspent_usd": C - staged}
```

With the invented release probabilities, expected outlay is $75.2m rather than $96m, leaving $20.8m in expected unspent funds. This is not yet a calculation of net option value. Staging is preferable only if the value of learning and redeployment exceeds any delay, administrative cost, funding uncertainty and foregone impact; useful milestones must also reveal information about success.

### Step 3

Step three is the memo itself: the three defensible answers, the crux table, and the reversal condition. Each answer is one a competent evaluator could give from the same evidence.

Table 127: Three defensible answers to the same memo. A1 and A3 differ not in risk tolerance but in what they believe a larger grant would buy that a smaller one would not.

| **Answer** | **Recommendation** | **Decisive consideration** | What would flip it |
|---|---|---|---|
| **A1** | Fund at the $96m tier, in three tranches, conditioned on independent financial audit and adoption of conflict-of-interest and harassment policies before tranche two | The illustrative break-even is 3.2 basis points at a payoff of 100 million life-equivalents; funding requires a defensible belief that the grant exceeds it and compares well with alternatives. P01 supports examining financial controls, governance, conflicts and harassment policies. | Evidence that the expansion rate exceeds the organization's demonstrated management capacity, or that a named alternative grantee has a materially higher marginal effect per dollar |

| **Answer** | **Recommendation** | **Decisive consideration** | What would flip it |
|---|---|---|---|
| **A2** | Fund at the $60m middle tier, hold $36m as an option, and re-underwrite in 12 months against hiring and research output | Proportionate due diligence (the role brief's own phrase): at $96m the evaluator cannot independently assess the management quality of an 80-person expansion, so fund at a level where the organization's own execution evidence is within reach of the next review | Evidence that the marginal research output per dollar is roughly constant across the $60m-$96m range, which would make the option valueless |
| **A3** | Decline the scale-up; recommend a $2-5m twelve-month planning grant and a reapplication with governance remediated and placement claims verified | D04's project-risk criterion plus P04's habit of asking early what evidence would change the decision: here the unverified placement claims and the single-founder dual role are the highest-likelihood failure modes, and a small grant buys the evidence a large one assumes | Independent verification of the placements and a named second executive with authority over research, either of which removes the premise of the decline |

Crux. The recommendation's sign turns on a single premise: whether the marginal $96m raises the probability of a good outcome by more than the bar re-

quires. Under the invented parameters that threshold is about three parts in ten thousand for a success worth $10^8$ life-equivalents. Almost nobody can estimate a quantity that small, and the honest memo says so rather than manufacturing the estimate. The second crux, which is easier to argue about and therefore more useful, is whether the marginal research output per dollar is roughly constant across the funding range; if it falls off sharply above the current size, the middle tier dominates regardless of the first crux.

Reversal condition, stated so that it can be checked rather than felt. Decline or drop to the small tier if either of the following holds: (i) the grant's marginal effect on the good-outcome probability is judged below $3.2 \times 10^{-4}$ under a decisive value of $10^8$ life-equivalents, or (ii) a named alternative use of the same money has a strictly higher marginal effect per dollar. Condition (ii) is the one the evaluator can actually investigate, and it is the one the memo should lead with.

Confidence. High confidence that the budget is legible and the unit costs are in range (support: direct comparison with P01's stated funding behaviour; a shallow but checkable assessment). Moderate confidence in the conditional-grant recommendation (support: a simple argument that does not depend on the invented parameters, combined with a source-stated best-practice list). Low confidence in any numerical estimate of the marginal probability (support: a general intuition; D01 would require that this be marked as such rather than presented as a modelled quantity).

## Sub-Question 3: Memo 2 - sub-$10k rapid grant: the evaluator's own cost structure is part of the answer

Invented premise. Lucia Ferrer requests $7,500 from a rapid-grants program to run four reading-group cycles at a Spanish university over two semesters. Budget as submitted: venue $1,200; tooling and subscriptions $300; speaker and facilitator honoraria $2,000; travel $1,500; part-time coordination stipend $2,500. Evidence of traction from one pilot cycle: 18 unique attendees, three of whom subsequently enrolled in an AI safety course. The program's published terms state grants of $50 to $10,000, average decisions within a day, nine in ten within a week, payment upfront by default, and that stipends and living expenses are not the usual scope but that applicants should make the case.

The comparators here are unusually concrete and they do most of the work. P13 reports a national meetup series funded at $4,800, a weekly evaluation reading-group organizer funded at $3,000, alongside an aspiration to scale the broader reading-group program from roughly 120 to 300 weekly attendees, city meetups at $1,000, facilitator stipends at $4,200, conference travel at $5,000, and $8,000 to validate a startup. P14, the earlier version of the same program, states that most grants fall between $50 and $1,500, that higher amounts require evidence of initial traction, and that compensation for the applicant's time is not funded. P14 also applies a high bar to conference travel: fund it when the applicant is

presenting or leading a session, when the conference is directly relevant, when other funding routes are exhausted, and when there is a specific deliverable beyond networking. The earlier P14 restrictions must not be presented as current D03 requirements; current guidance permits travel and invites a stronger case for stipends.

**The non-trivial move in this memo is not the fund-or-decline call. It is recognising that at this size the evaluator's own hours are a material share of the grant, so the correct home for the proposal is partly determined by the evaluator's cost structure and mandate rather than by the proposal's quality. The same proposal has different correct answers at BlueDot, at LTFF, and at Coefficient Giving, and the corpus supports each of those answers.**

### Step 1

Step one computes cost per outcome and places it against the comparators the source itself reports, being explicit that some comparator denominators are vague and must be handled as ranges.

```
request = 7_500
lines = {"venue": 1_200, "tooling": 300, "honoraria": 2_000,
    "travel": 1_500, "coordination_stipend": 2_500}
assert sum(lines.values()) == request

# Source comparators (P13) and the traction evidence from the
    invented pilot.
attendees_pilot = 18
cycles = 4
implied_attendees = attendees_pilot * cycles  # attendance
    opportunities; unique people may overlap
pipeline_entries = 3 * cycles  # hypothetical extrapolation of the
    pilot rate to four cycles

print(f"cost per discounted outcome line: ${request:,}")
print(f"cost per attendee at pilot-derived turnout:
    ${request/implied_attendees:,.0f}")
print(f"cost per pipeline entry at pilot-derived rate:
    ${request/pipeline_entries:,.0f}")

# The comparator with the most informative denominator is P13's
    national meetup series.
comparators = [
    {"label": "AI Safety Poland national meetups (P13)", "usd": 4_800,
     "attendance_lo": 200, "attendance_hi": 400, "note": "200-400 is
    an invented illustration of the word hundreds, not a
    source-reported bound"},
    {"label": "weekly evals reading group (P13)", "usd": 3_000,
     "attendance_lo": 120, "attendance_hi": 300, "note": "120 weekly
```

```
        attendees to a target of 300 across the broader reading-group
        program; not an achieved total attributable to this grant"},
]
for c in comparators:
    lo = c["usd"] / c["attendance_hi"]
    hi = c["usd"] / c["attendance_lo"]
    print(f'{c["label"]}: ${lo:,.1f}-${hi:,.1f} per attendee
        ({c["note"]})')

# Scope screen against the programme's own published terms.
scope = {
    "venue":          "in scope (D03 names events and venue costs)",
    "tooling":        "in scope (D03 names project tooling; P14 funds
        project-specific software)",
    "honoraria":      "precedent exists (P13 funded $4,200 of
        facilitator stipends)",
    "travel":         "current D03 allows conference, collaboration
        and fieldwork travel when its relevance is explained; P14
        describes an older, stricter test",
    "coordination_stipend": "current D03 invites a stronger
        high-impact case; P14 prohibited compensation under older
        guidance",
}
in_scope_default = lines["venue"] + lines["tooling"] +
    lines["honoraria"]
print("default-scope total: $", f"{in_scope_default:,}")
print("plus travel if its work-related value is justified under
    current guidance: $", f"{in_scope_default + lines['travel']:,}")

output_state["memo_2_unit_costs"] = {"per_attendee": request /
    implied_attendees,
                                     "per_pipeline_entry": request /
    pipeline_entries,
                                     "default_scope_usd":
    in_scope_default}
output_state["memo_2_scope_screen"] = scope
```

The comparator cost ratios are not like-for-like unit costs. P13 does not supply a bounded 200–400 attendance count for Poland, and 300 weekly reading-group attendees is a target rather than an achieved denominator attributable to one grant. In the invented pilot, multiplying 18 attendees by four cycles yields 72 attendance opportunities, not necessarily 72 unique people. Current D03 allows travel and invites stronger stipend cases; it does not reject those two lines categorically. The $2,500 stipend is one third of the request.

### Step 2

Step two is the part most evaluators skip: the cost of the evaluation itself. If a funder's loaded cost per grantmaker hour is $r$, the memo takes $h$ hours, and the

target maximum share of grant size spent on evaluation is $\alpha$, then the minimum grant the funder can rationally process is:

$$g_{\min} = \frac{h\,r}{\alpha}$$

```
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt
import numpy as np

def g_min(hours, loaded_rate, alpha):
    return hours * loaded_rate / alpha

# All hours, loaded rates and the 10% constraint below are invented scenario inputs, not measured funder operating costs.
scenarios = [
    ("BlueDot rapid grant (self-serve, ~1 day decision)", 0.4, 60, 0.10),
    ("ACX Grants (volunteer expert evaluators)",           2.0, 80, 0.10),
    ("Manifund regrantor (part-time, own pot)",            3.0, 90, 0.10),
    ("LTFF-style small fund (full-time team)",             8.0, 110, 0.10),
    ("Coefficient Giving (full investigation)",           40.0, 110, 0.10),
]
for label, h, r, a in scenarios:
    print(f'{label:52s} minimum efficient grant: ${g_min(h, r, a):>10,.0f}')

# Where does the invented $7,500 request sit against each structure?
request = 7_500
for label, h, r, a in scenarios:
    evaluation_cost = h * r
    share = evaluation_cost / request
    verdict = "processable" if share <= a else "below this funder's efficient size"
    print(f'{label:52s} evaluation cost ${evaluation_cost:>7,.0f} = {share:6.1%} of the grant -> {verdict}')

fig, ax = plt.subplots(figsize=(9.5, 5.5))
hs = np.linspace(0.2, 45, 200)
for r, style in [(60, ":"), (90, "--"), (110, "-")]:
    ax.loglog(hs, g_min(hs, r, 0.10) / 1000.0, style, color="#2b6cb0",
              label=f"loaded rate ${r}/hour, evaluation share <= 10%")
for y, lbl, col in [(0.65, "BlueDot implied mean (P13)", "#2f855a"),
                    (10.0, "BlueDot ceiling (D03)", "#2f855a"),
                    (24.0, "ACX planned mean (P16)", "#6b46c1"),
```

```
                    (42.7, "LTFF observed mean (P10)", "#6b46c1"),
                    (200.0, "D04 first-grant band floor", "#c53030")]:
    ax.axhline(y, ls="-.", lw=1, color=col, alpha=0.7)
    ax.annotate(lbl, (hs[0], y), xytext=(4, 3), textcoords="offset
    points", fontsize=8, color=col)
ax.scatter([1.0], [7.5], marker="*", s=170, color="#c53030", zorder=6,
           label="invented request: $7.5k after ~1 hour")
ax.set_xlabel("grantmaker hours per decision (log scale)")
ax.set_ylabel("minimum efficient grant, $ thousands (log scale)")
ax.set_title("Evaluation cost sets a floor on grant size, and the
    floor differs by funder")
ax.legend(fontsize=8, loc="upper left")
ax.grid(alpha=0.3, which="both")
fig.tight_layout()
fig.savefig(_generated_image_path_, dpi=160)
plt.close(fig)

output_state["memo_2_eval_cost"] = {label: {"hours": h, "rate": r,
    "eval_cost": h * r,
                                            "share_of_7500": h * r /
    request}
                                    for label, h, r, a in scenarios}
```

Under the invented hours, rates and 10% overhead-share rule, some funders' hypothetical processes fit this grant size and others do not. That illustrates one mechanism for comparative advantage; it does not prove a request irrational to process or establish actual institutional costs. D04 and P10 motivate proportionality, while exceptional impact or a cheaper review process can justify an exception.

### Step 3

Step three: the memo. Note that all three answers accept the pilot evidence and differ on scope and on who should pay.

Table 128: B1 and B2 agree on the merits and differ on scope; B3 denies the mandate rather than the proposal. Only B3 is a statement about the evaluator, not the application.

| **Answer** | **Recommendation** | **Decisive consideration** | What would flip it |
|---|---|---|---|
| **B1** | Fund $7,500 in full, paid upfront | Under the invented rapid-grant assumptions, $24 of evaluation cost is 0.32% of the $7,500 request, so the request is above the illustrative $240 processing floor. Traction and P13's person-level funding precedents support assessing the full request, subject to incremental value and current scope. | Evidence that the coordinator's time is not the binding constraint, in which case the stipend is a transfer rather than a production input |
| **B2** | Fund $3,500 for venue, tooling and honoraria; consider up to $5,000 if the travel's work-related value is justified under current D03 guidance; request the stronger case needed for the coordination stipend. | Current guidance allows travel and asks for a stronger stipend case. A smaller initial recommendation can be justified by missing evidence about incremental value, not by importing the older P14 exclusions. | A statement that the cycles will not run at all without a stipend, which makes the stipend the marginal input rather than a cost line |

| Answer | Recommendation | Decisive consideration | What would flip it |
|---|---|---|---|
| **B3** | Decline and route: the applicant is eligible for an existing rapid-grant program and for a small fund whose evaluation cost is proportionate | Comparative advantage: at this size a full investigation consumes a share of the grant that would be better spent on the work, and the corpus names alternative funders who serve this band | Evidence that no other funder in the band will take the request, in which case the evaluation cost is the price of the counterfactual, not waste |

Crux. The recommendation turns on whether the coordination stipend is a production input or a transfer. If the four cycles happen only because someone is paid to coordinate them, the stipend is the highest-leverage line in the budget and B2 is wrong. If the applicant would run them anyway and the stipend substitutes for income, the line is a transfer, and a funder with a stated policy against funding compensation for time should not be the payer. This is exactly the distinction P13 draws when it says it started treating rapid grants as small bets on talented people, including cases where the program approached people directly and pitched them on applying for work they were already doing.

Reversal condition. Flip from fund to decline-and-route if the marginal value of the grant depends on the coordination line and the funder's own mandate forbids funding it, because the counterfactual then belongs to a different funder. Flip from decline-and-route to fund if the applicant demonstrates that no funder in the band will take a $7,500 request without a full application, which is precisely the friction P10 says reliable early capital is meant to remove.

Confidence. High confidence that current D03 allows travel and invites stronger stipend cases. Low confidence in unit-cost comparisons based on vague, projected or non-matching attendance denominators, and in whether the activities would occur without the stipend. Those are unresolved inputs, not grounds for treating historical exclusions as current policy.

## Sub-Question 4: Memo 3 - policy and field-building: pay for the cheapest relative improvement, not the most persuasive stage

Invented premise. The Centre for AI Policy Feasibility requests $1.2m over two years, which is $600k per year and sits inside D04's first-grant band. The

team is two senior researchers (one a former ministry official with six years in government), one part-time communications lead, and one research assistant. There are four work packages: WP1, three jurisdictional pipeline maps; WP2, two response-development papers; WP3, twelve policymaker meetings and one workshop with 25 participants; WP4, one policy brief pitched at adoption. Claimed track record: one earlier report cited in a parliamentary written answer, which the memo cannot verify. Claimed theory of change: research produces responses, responses attract a coalition, a coalition produces a drafted instrument, the instrument is adopted, and adoption is implemented and effective.

N14 is the governing comparator and it is unusually specific. It reports that the most common reason proposals were rejected was political infeasibility; that initiatives most likely to progress were those incurring little or no financial or political cost to the politician; that two well-designed instruments failed because they presupposed an expert consensus that did not yet exist and because the institutions could not do original research; that the field lacked coalition-building expertise; and that the successes were softer than hoped, namely a report widely circulated through several departments and a chapter in the government chief scientist's annual report. N14 also states the operative conclusion directly: the community should build coalitions around the problem and identify solutions, and should not yet shift attention to the later stages.

*The evaluator's first instinct on policy work is to say that it is hard and slow. That instinct is not wrong, but it is not an evaluation. The evaluation is to locate the proposal in a pipeline, ask whether the pipeline is ready for work at that location, and then ask which location the marginal dollar moves most cheaply.*

**Step 1**

Step one models the pipeline. Let the end-to-end probability be the product of per-stage probabilities, $p = \prod_i p_i$. A relative improvement $r$ at stage $i$ multiplies the end-to-end probability by $(1+r)$, so the value gained is proportional to $k\,p\,r$ and is therefore the same at every stage for a given $r$.

$$\Delta V_i \;=\; k\,p\,\frac{\partial \ln p}{\partial \ln p_i}\,r_i \;=$$

$$k\,p\,r_i \qquad \text{(multiplicative pipeline)}$$

**Within this multiplicative model, equal relative improvements at different stages have equal first-order value, so compare their marginal cost and feasibility. The least-funded stage is not automatically cheapest, and the adoption stage is not automatically the wrong choice. Those claims require evidence about the available interventions and cost curves.**

```python
import numpy as np

# Invented per-stage probabilities for the policy pipeline.
p = {"response": 0.45, "coalition": 0.35, "draft": 0.30, "adopt": 0.25, "implement": 0.55}
end_to_end = float(np.prod(list(p.values())))
print(f"end-to-end probability of policy influence: {end_to_end:.4f} ({end_to_end:.2%})")

# Invented cost, in dollars, to buy a 20% RELATIVE improvement at each stage.
cost_for_20pct = {"response": 300_000, "coalition": 450_000,
                  "draft": 1_100_000, "adopt": 4_000_000, "implement": 2_600_000}

print("\nCheapest relative improvement per dollar:")
ranked = sorted(cost_for_20pct, key=lambda s: cost_for_20pct[s])
for s in ranked:
    print(f'  {s:10s} ${cost_for_20pct[s]:>10,} per +20% relative improvement'
          f'  (= ${cost_for_20pct[s]/0.20:>11,.0f} per unit of relative improvement)')

# The invented proposal's own allocation, mapped to the stage each package moves.
allocation = {"response": 450_000, "coalition": 250_000, "adopt": 300_000, "implement": 0, "draft": 200_000}
print("\nProposal allocation by stage:", allocation, "total:", f"${sum(allocation.values()):,}")

# A linear approximation to the per-stage production of relative improvement.
# Declared as an approximation: real cost curves are convex in r, and the convexity
# is what caps how much relative improvement a fixed budget can buy at a single stage.
def relative_gain(stage, usd):
    return usd / cost_for_20pct[stage] * 0.20

base = sum(relative_gain(s, v) for s, v in allocation.items())
realloc = dict(allocation)
shift = 300_000
realloc["adopt"] -= shift
realloc["response"] += shift
better = sum(relative_gain(s, v) for s, v in realloc.items())
print(f"\nrelative improvement, as submitted:        {base:6.1%}")
print(f"relative improvement, after reallocating $300k from adopt to response: {better:6.1%}")
print(f"ratio: {better/base:.2f}x, at identical cost")
```

```

output_state["memo_3_pipeline"] = {"stage_probabilities": p,
    "end_to_end": end_to_end,
                                   "cost_for_20pct": cost_for_20pct,
                                   "relative_gain_as_submitted": base,
                                   "relative_gain_reallocated": better}
```

*Under the invented parameters, reallocating $300k raises the linear approximation to relative improvement from 46.25% to 64.75%: about 1.40 times the original gain, or a 40.0% increase, at the same $1.2m cost. This does not double expected value. The ranking depends on the assumed costs of stage improvements and on the model; N14 motivates examining neglected stages but does not establish these numerical cost curves.*

One caveat the memo must state: real cost curves are convex in the relative improvement a stage can absorb. Doubling response-stage output from a very low base may be cheap, but quadrupling it is not, and the linear model overstates what a single reallocation buys. The conclusion that survives convexity is weaker but still action-guiding: do not fund the stage whose pitch is most persuasive; fund the stage where the next unit of relative improvement is cheapest, and check that the stage can absorb it.

### Step 2

Step two applies N14's political-feasibility screen as a multiplier rather than a veto. The screen's four questions are: does the proposal impose financial cost on the government; does it impose political cost on identifiable actors; does it presuppose an expert consensus that does not exist; and is there a champion inside the institution that would have to act.

```
import numpy as np

# Invented feasibility factors, one per N14-style screen question,
    applied per work package.
# Factors are multiplicative: they answer 'what fraction of the
    expected value survives this screen'.
screen = {
    "WP1 jurisdictional pipeline maps": {"financial_cost_to_gov":
    0.95, "political_cost": 0.95,
                                         "presupposes_consensus":
    0.90, "champion_available": 0.85},
    "WP2 response-development papers":  {"financial_cost_to_gov":
    0.95, "political_cost": 0.95,
                                         "presupposes_consensus":
    0.80, "champion_available": 0.80},
    "WP3 policymaker meetings":        {"financial_cost_to_gov":
    0.80, "political_cost": 0.55,
                                         "presupposes_consensus":
    0.45, "champion_available": 0.60},
```

```
    "WP4 policy brief targeted at adoption":
    {"financial_cost_to_gov": 0.75, "political_cost": 0.45,
                                        "presupposes_consensus":
    0.35, "champion_available": 0.55},
}

for wp, factors in screen.items():
    m = float(np.prod(list(factors.values())))
    print(f"{wp:42s} feasibility multiplier: {m:.3f}")

m3 = float(np.prod(list(screen["WP3 policymaker meetings"].values())))
m1 = float(np.prod(list(screen["WP1 jurisdictional pipeline
    maps"].values())))
print(f"\nratio of feasibility multipliers, WP1 over WP3:
    {m1/m3:.2f}x")

# Combined screen: multiply the linear relative-improvement gain by
    the feasibility multiplier
# for each work package, so that the memo's ordering reflects both
    stage cost and stage feasibility.
package_usd = {"WP1": 450_000, "WP2": 250_000, "WP3": 300_000, "WP4":
    200_000}
stage_of = {"WP1": "response", "WP2": "coalition", "WP3": "adopt",
    "WP4": "adopt"}
multiplier = {"WP1": m1, "WP2": float(np.prod(list(screen["WP2
    response-development papers"].values()))),
              "WP3": m3, "WP4": float(np.prod(list(screen["WP4 policy
    brief targeted at adoption"].values())))}
cost_for_20pct = {"response": 300_000, "coalition": 450_000, "draft":
    1_100_000,
                  "adopt": 4_000_000, "implement": 2_600_000}

rows = []
for wp in package_usd:
    r = package_usd[wp] / cost_for_20pct[stage_of[wp]] * 0.20
    rows.append((wp, stage_of[wp], package_usd[wp], r,
    multiplier[wp], r * multiplier[wp]))
rows.sort(key=lambda t: -t[5])
print("\nwork package | stage | usd | raw relative gain | feasibility
    | screened gain")
for wp, st, usd, r, m, sr in rows:
    print(f"{wp} | {st:9s} | ${usd:>7,} | {r:6.1%} | {m:5.3f} |
    {sr:6.1%}")
print(f"\nscreened total: {sum(t[5] for t in rows):.1%} versus
    {sum(t[3] for t in rows):.1%} unscreened")

output_state["memo_3_feasibility"] = {"screen": screen,
    "screened_rows": rows,
                                      "screened_total": sum(t[5] for t
    in rows)}
```

The screen does not change the direction of the reallocation; it strengthens it. WP3 and WP4 are the packages that load most heavily on the political-cost and consensus questions, and they are the packages whose screened gain collapses. N14's own observation that the field lacked coalition-building expertise cuts the other way, however, and that is the strongest argument for keeping some WP3: if the expertise does not exist elsewhere, paying to build it has option value even when the immediate policy output is unlikely.

**Step 3**

Step three: the memo, with the three answers, the crux, and the reversal condition.

Table 129: C1 and C2 share the pipeline analysis and differ on whether an unfunded stage should be seeded. C3 is the scope answer rather than the stage answer.

| **Answer** | **Recommendation** | **Decisive consideration** | What would flip it |
|---|---|---|---|
| **C1** | Fund $1.2m over two years, conditional on reallocating roughly $300k from advocacy to response development and coalition mapping, with a milestone that names one decision-maker who received and used an output | Relative improvements are equally valuable at every pipeline stage under a multiplicative model, so the money should buy the cheapest relative improvement; the screen independently penalises the advocacy packages | Evidence that WP3's meetings are the binding input, for example a named official who has already requested a specific response |

| Answer | Recommendation | Decisive consideration | What would flip it |
|---|---|---|---|
| **C2** | Fund $900k over two years as a research-only project, drop the advocacy package entirely, and require the team to publish the pipeline maps | N14's finding that policy work done before a coalition exists around the problem is premature; the funder should not pay for a stage the pipeline is not ready for, even at a good price | A demonstrated coalition around the problem in at least one jurisdiction, verifiable through third parties rather than self-report |
| **C3** | Decline and re-scope toward a single jurisdiction with a named institutional partner and a deliverable the partner has committed to use | The unverified citation claim plus the breadth of three jurisdictions inside a two-year, four-person team; D04's scale preference rewards a tighter scope before a larger one | Verification of the citation claim together with a letter of intent from one jurisdiction, which removes both the evidence gap and the breadth problem |

**Crux. The sign of the recommendation turns on whether the pipeline stage probability that the money moves is the stage that is currently binding. If adoption is genuinely gated on the quality of responses that do not yet exist, then funding response development is the high-value move and the advocacy package is premature. If adoption is instead gated on access and relationships that the team already has, then WP3 is the binding input and the arithmetic above is irrelevant to the decision. The memo cannot settle this from the documents; it can only name the question and say who could answer it.**

Reversal condition. Flip from C1 to C2 if the response stage is already well supplied in at least two of the three jurisdictions, because then the next unit of relative improvement there is expensive and the cheapness argument inverts. Flip from C1 to C3 if the funded work cannot name a specific decision-maker audience, since D04 explicitly asks for a realistic scope and, for research, an audience and often a target venue.

Confidence. High confidence that the multiplicative-stage argument is the right frame, because it re-derives N14's conclusion from the structure of the problem rather than from its examples. Moderate confidence in the feasibility multipliers, which are invented. Low confidence in the claim that the response stage is undersupplied in these particular jurisdictions; support for that claim is currently a general impression, and D01 requires that it be labelled as one.

## Sub-Question 5: Sensitivity: the same three memos under short versus long timelines and high versus low risk tolerance

This sub-question runs the required exercise: recompute all three recommendations under two axes. The timeline axis runs long, medium, short. The risk-posture axis runs low tolerance and high tolerance. The claim to be tested is that the two axes do different work. Risk posture should move the threshold a grant must clear, and therefore the fund-decline line. Timelines should change what is being maximised, and therefore the ranking between opportunities.

That claim is an inference of the author's, not a source-supported finding. It is offered as a testable proposition with a decision consequence: if true, an evaluator can absorb disagreement about risk tolerance by adjusting a number, but cannot absorb disagreement about timelines without changing which proposals they prefer. The corpus supplies grounds for suspicion in both directions. P01 and P06 report shortened timelines and P08 records a 2016 estimate of a 10% chance of transformative AI by 2036, so institutional timelines have moved. S11 states that timelines might be medium or broad, and P03 conditions its entire analysis on insiders being roughly right.

### Step 1

Step one encodes the three opportunity profiles on three dimensions, defines two weight regimes for timelines and a threshold shift for risk posture, and records the decision in each cell.

```
import numpy as np

# Invented profile scores in [0,1] for each opportunity, read off the
    three memos.
profiles = {
    "A large scale-up":   {"durable_capacity": 0.85,
    "near_term_default": 0.55, "speed_to_impact": 0.20},
    "B rapid grant":      {"durable_capacity": 0.30,
    "near_term_default": 0.45, "speed_to_impact": 0.95},
    "C policy/field":     {"durable_capacity": 0.65,
    "near_term_default": 0.60, "speed_to_impact": 0.40},
}

dims = ["durable_capacity", "near_term_default", "speed_to_impact"]
```

```

timeline_weights = {
    "long":   {"durable_capacity": 0.60, "near_term_default": 0.25,
    "speed_to_impact": 0.15},
    "medium": {"durable_capacity": 0.45, "near_term_default": 0.35,
    "speed_to_impact": 0.20},
    "short":  {"durable_capacity": 0.30, "near_term_default": 0.50,
    "speed_to_impact": 0.20},
}
# Risk posture shifts the threshold, not the weights.
threshold = {"low_tolerance": 0.62, "high_tolerance": 0.48}

rows = []
for tl, w in timeline_weights.items():
    for rp, tau in threshold.items():
        scores = {k: sum(w[d] * v[d] for d in dims) for k, v in
    profiles.items()}
        funded = sorted([k for k, s in scores.items() if s >= tau],
    key=lambda k: -scores[k])
        ranking = sorted(scores, key=lambda k: -scores[k])
        rows.append({"timeline": tl, "risk": rp, "scores":
    dict(scores),
                     "ranking": ranking, "funded": funded})

for r in rows:
    print(f'{r["timeline"]:6s} | {r["risk"]:14s} | ranking: {" >
    ".join(r["ranking"])}'
          f' | funded: {r["funded"] if r["funded"] else "none"}')

rank_changes = {}
for tl in timeline_weights:
    long_like = [r for r in rows if r["timeline"] == tl]
    for r in long_like:
        rank_changes[(tl, r["risk"])] = r["ranking"][0]
print("\ntop-ranked opportunity by cell:", rank_changes)

output_state["sensitivity_grid"] = rows
```

Under these invented profiles, A leads under long and medium timelines, while C narrowly leads under short timelines. B never ranks first. At the stricter 0.62 threshold, only A under long timelines passes; B fails all three regimes. At the 0.48 threshold, A and C pass in every regime, and B passes only under medium and short timelines. Changing the threshold changes eligibility without changing the ranking within a timeline regime.

*The rapid grant's scores are 0.435, 0.4825 and 0.505 across long, medium and short timelines. This toy grid does not establish that it is robustly fundable across worldviews. It illustrates how eligibility depends on both the selected weights and the threshold; the dealflow and regranting arguments in P11 and*

*P12 require separate assessment.*

### Step 2

Step two computes the flip points and, more importantly, the crossover between one large grant and a portfolio of small ones. The crossover is the sharpest version of the disagreement in the corpus between centralised large funders (P01, P06) and dispersed regranting (P11, P12).

$$\frac{p_A V_A}{C_A} = \frac{p_B V_B}{C_B} \quad \Longleftrightarrow \quad V_B = \frac{C_B}{C_A} \cdot \frac{p_A}{p_B} V_A .$$

```python
import numpy as np
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt

C_A, C_B = 96_000_000.0, 7_500.0            # invented: one large grant, one rapid grant
p_A, p_B = 0.030, 0.004                      # invented marginal probabilities

# Value per hit at which the two options have equal expected value per dollar.
V_B_over_V_A = (C_B / C_A) * (p_A / p_B)
print(f"crossover: a rapid grant's hit must be worth {V_B_over_V_A:.6%} of the large grant's hit")

# Equivalent statement in dollars, using an invented large-grant hit value.
V_A = 1e8
V_B_star = V_B_over_V_A * V_A
print(f"with V_A = {V_A:,.0f} life-equivalents, the crossover hit value is {V_B_star:,.0f}")
print(f"i.e. a rapid-grant hit worth about {V_B_star/V_A*100:.4f}% of the large one")

# Portfolio arithmetic at the rapid-grant size, per the earlier phase's power-law model.
c_B = C_B
B = 2_000_000.0
n = int(B / c_B)
p_hit = 0.01
print(f"\nwith ${B:,.0f} the funder buys {n} rapid grants at ${c_B:,.0f}")
print(f"probability of at least one hit at p = {p_hit}: {1-(1-p_hit)**n:.3f}")
```

```
26
27 # Flip points: the value of p_A relative to p_B at which the large
       grant wins on EV per dollar.
28 for hit_multiplier in [1, 2, 5, 10, 50]:
29     p_A_star = p_B * hit_multiplier * (C_A / C_B) * (V_B_star / V_A)
30     feasible = "feasible probability" if p_A_star <= 1 else
       "impossible: threshold exceeds 1"
31     print(f"if V_B is {hit_multiplier:>3d} times the reference
       V_B_star, break-even p_A = {p_A_star:.3f} ({feasible})")
32
33 fig, axes = plt.subplots(1, 2, figsize=(12.5, 5.2))
34
35 ax = axes[0]
36 ratios = np.logspace(-3, 4, 200)
37 ax.loglog(ratios, (C_B / C_A) * ratios * V_A, color="#c53030")
38 ax.axhline(V_A, ls=":", color="#718096")
39 ax.annotate("large-grant hit value", (ratios[0], V_A), xytext=(4, 4),
       textcoords="offset points", fontsize=8, color="#718096")
40 ax.set_xlabel("ratio p_A / p_B")
41 ax.set_ylabel("crossover hit value for the rapid grant,
       life-equivalents")
42 ax.set_title("Crossover: when one rapid grant matches\none large
       grant in expected value per dollar")
43 ax.grid(alpha=0.3, which="both")
44
45 ax = axes[1]
46 labels = ["long", "medium", "short"]
47 dim_labels = ["durable capacity", "near-term default", "speed to
       impact"]
48 weights = np.array([[0.60, 0.25, 0.15], [0.45, 0.35, 0.20], [0.30,
       0.50, 0.20]])
49 bottom = np.zeros(3)
50 colours = ["#2b6cb0", "#dd6b20", "#38a169"]
51 for j, dl in enumerate(dim_labels):
52     ax.bar(labels, weights[:, j], bottom=bottom, label=dl,
       color=colours[j], alpha=0.85)
53     bottom += weights[:, j]
54 ax.set_ylabel("weight on each objective")
55 ax.set_title("Timelines change what is being maximised;\nrisk posture
       changes only the threshold")
56 ax.legend(fontsize=8)
57 ax.grid(axis="y", alpha=0.3)
58
59 fig.tight_layout()
60 fig.savefig(_generated_image_path_, dpi=160)
61 plt.close(fig)
62
63 output_state["crossover"] = {"V_B_over_V_A": V_B_over_V_A,
       "V_B_star_at_V_A_1e8": V_B_star,
64                              "portfolio_n": n,
```

```
        "portfolio_p_at_least_one": 1 - (1 - p_hit) ** n}
```

**With the invented probabilities $p_A = 0.030$ and $p_B = 0.004$, equal expected value per dollar occurs when a rapid-grant hit is worth 0.05859375% of a large-grant hit, about one part in 1,707. A larger rapid-grant payoff makes B better on this metric; a smaller one makes A better, holding the other inputs fixed. The calculation does not make expected value per dollar uninformative: it shows why the joint estimates of marginal probability and payoff matter. P11's $143k Timaeus regrant followed by a $160m grant is one reported funding trajectory, not an estimate of a success probability or final impact.**

A second caution belongs here. Any argument that a portfolio of small grants dominates by diversification must specify what the hits have in common. If the portfolio's grants share a failure mode, such as all depending on the same two or three community gatekeepers, then the independence assumption behind $1 - (1 - p)^n$ is false and the diversification argument weakens sharply. P12's point that the community has historically run on mutual trust that probably will not scale is a reason to expect correlated failure, not independent failure.

### Step 3

Step three records the answer table for the six cells, then the one-paragraph dissent that the exercise requires.

Table 130: Risk posture and timelines move the recommendations, but three components are invariant across the whole table. Those invariants are the parts of the answer that survive an interviewer's worldview.

| Regime | Memo 1 change | Memo 2 change | Memo 3 change |
|---|---|---|---|
| Long timelines, low tolerance | A2 (middle tier) becomes the default; durable capacity outweighs speed and the option is worth having | B2 (de-scoped), unchanged; the argument is scope, not urgency | C2 (research-only) becomes the default; coalition work can wait for the field to mature |

| Regime | Memo 1 change | Memo 2 change | Memo 3 change |
|---|---|---|---|
| Short timelines, high tolerance | A1 (full tier, staged) becomes the default; the cost of under-building now rises | B1 (full request) becomes defensible; speed and person-level bets are the point | C1 (reallocated, including advocacy) becomes the default; the window argument strengthens the adoption stage |
| Regime-invariant component | The governance condition never disappears: P01's best-practice list is not a function of timelines | The scope screen never disappears: it comes from the funder's published terms, not from the worldview | The stage analysis never disappears: the cheapest relative improvement is a property of the pipeline |

Dissent paragraph, required by the exercise. The strongest alternative answer to this whole drill is that the sensitivity analysis is a category error, because the recommendations never depended on the worldview in the first place: the governance condition on Memo 1, the scope screen on Memo 2, and the stage analysis on Memo 3 are all derived from the funder's published terms, the grantee's own budget, and the structure of the policy pipeline, and none of those three inputs changes when a forecast about timelines changes. On that view, the drill has demonstrated only that a competent evaluator arrives at the same recommendation by three different routes, and the exercise's value lies in showing how much of the answer is worldview-invariant rather than in producing different answers under different worldviews. The dissent is not idle: it predicts that a mock interview which varies stated risk tolerance will produce more superficial variation in answers than one which varies the stated state of the field, and it can be tested by running the same proposal past two evaluators with different stated worldviews and comparing where the disagreement actually lands.

## Sub-Question 6: Alternative defensible answers, reversal conditions, and final calibration

The role brief asks for concise, calibrated, reasoning-transparent communication. This sub-question converts the three memos into that register: one paragraph per memo giving the strongest alternative answer, one numeric reversal condition per memo that a decision-maker could check, and one calibration sentence per memo that names the confidence and the kind of support behind it, in D01's vocabulary.

*The discipline that matters most here is refusing to claim more precision than the evidence supports, and refusing to claim less. The corpus is generous with operational facts and stingy with outcome probabilities. A memo that reports a probability it cannot support has failed the reasoning-transparency test even if the number turns out to be right; a memo that hides behind 'it depends' has failed the concision test even if every hedge is warranted.*

### Step 1

Step one assembles the reversal thresholds from the three memos into one table, expressed as the quantity a decision-maker would have to observe to overturn each recommendation.

```
# Summarize quantities already derived in the exercise; do not invent reversal evidence.
C, b, V = 96_000_000.0, 1.0 / 3_000.0, 100_000_000.0
p_hypothetical = 0.00030
p_break_even = b * C / V
policy_gain = 0.6474747474747475  # Step 1 linear reallocation output
reversals = {
    "Memo 1: governance + staged funding": {
        "central": p_hypothetical, "reversal": p_break_even,
        "unit": "marginal probability for the full $96m at the stated benchmark and payoff",
        "status": "hypothetical probability is below break-even; no demonstrated safety margin",
        "flip_to": "reassess, reduce scope, or decline if expected impact does not clear the bar",
        "observable": "evidence about marginal impact, possible harms, and alternatives"},
    "Memo 2: de-scoped rapid grant": {
        "central": 3500, "reversal": 7500,
        "unit": "USD funding options; these are not a quantified decision threshold",
        "flip_to": "full request, or decline and route after assessing incremental value and current scope",
        "observable": "whether the coordination payment changes the work and is justified"},
    "Memo 3: reallocated policy grant": {
        "central": policy_gain, "reversal": None,
        "unit": "linear relative improvement per $1.2m under the invented allocation model",
        "flip_to": "reassess allocation or scope when evidence changes stage cost or feasibility",
        "observable": "jurisdiction-specific response-stage supply and marginal intervention costs"},
}
for name, row in reversals.items():
    print(name, row)
```

```
print(f"Memo 1: hypothetical p={p_hypothetical:.6f};
    break-even={p_break_even:.6f};
    ratio={p_hypothetical/p_break_even:.4f}")
print("No factor-of-six robustness follows from these assumptions.")
output_state["reversals"] = reversals
```

The invented probability 0.00030 is 93.75% of the earlier break-even 0.00032, so it falls short rather than providing a factor-of-six safety margin. No empirical central probability has been established. The rapid-grant funding tiers and the policy reallocation output are useful decision inputs, but neither supplies a numeric reversal threshold on its own.

### Step 2

The figure displays different comparison quantities for each memo. Memo 1 compares the hypothetical probability with its derived break-even; Memo 2 compares two funding options; Memo 3 shows a modeled gain whose reversal threshold remains unknown. Normalizing them does not make them comparable measures of robustness.

```
import matplotlib
matplotlib.use("Agg")
import matplotlib.pyplot as plt

# Normalize known comparison quantities within each memo; a missing
    threshold stays missing.
rows = [
    ("Memo 1: marginal probability", 0.00030 / 0.00032, 1.0),
    ("Memo 2: funding options (not thresholds)", 3500 / 7500, 1.0),
    ("Memo 3: modeled relative gain", 1.0, None),
]
fig, ax = plt.subplots(figsize=(10, 4.6))
for y, (label, first, second) in enumerate(rows):
    ax.plot([first], [y], marker="o", markersize=10, color="#2b6cb0")
    if second is not None:
        ax.plot([first, second], [y, y], lw=2, color="#a0aec0")
        ax.plot([second], [y], marker="X", markersize=10,
    color="#c53030")
    else:
        ax.annotate("reversal threshold not established", (first, y),
    xytext=(0, 15), textcoords="offset points", ha="center",
    fontsize=8)
ax.set_yticks(range(len(rows)))
ax.set_yticklabels([r[0] for r in rows], fontsize=9)
ax.set_xlim(0, 1.25)
ax.set_xlabel("Memo 1: fraction of break-even; Memo 2: fraction of
    full request; Memo 3: gain normalized to itself")
ax.set_title("Derived quantities and funding options; not a common
    robustness score")
```

```
ax.grid(axis="x", alpha=0.25)
fig.tight_layout()
fig.savefig(_generated_image_path_, dpi=160)
plt.close(fig)
output_state["calibration_rows"] = rows
```

The $3,500 and $7,500 amounts do not establish that Memo 2 is closest to a decision reversal. They are alternative grant sizes. Whether to include the stipend depends on its incremental value and the current funder's criteria, not on the distance between the two dollar amounts.

### Step 3

Step three writes the three calibration statements in D01's vocabulary, and then the closing note on what this drill cannot establish.

```
# Calibration register: claim, confidence, kind of support, and the
    honest residual.
register = [
    {"memo": "Memo 1", "claim": "the budget is legible and the unit
    costs are in range",
     "confidence": "high",
     "support": "careful comparison with a source-stated funding
    posture (P01), shallow but checkable",
     "residual": "no independent data on this organisation's
    execution"},
    {"memo": "Memo 1", "claim": "the marginal probability of a
    decisive effect exceeds the bar",
     "confidence": "low",
     "support": "a simple argument that does not depend on the
    invented parameters, plus a general intuition",
     "residual": "the quantity itself is unmeasurable from
    documents"},
    {"memo": "Memo 2", "claim": "current guidance allows travel and
    requests a stronger stipend case",
     "confidence": "high",
     "support": "direct reading of current D03, distinguished from
    historical P14 and expanded scope in P13",
     "residual": "whether the applicant can make the case those terms
    invite"},
    {"memo": "Memo 2", "claim": "the cycles would not happen without
    a stipend",
     "confidence": "low",
     "support": "a vague impression based on reading comparable grant
    reports",
     "residual": "a fact about a person, not a fact about a
    document"},
    {"memo": "Memo 3", "claim": "relative improvements are equally
    valuable at every pipeline stage",
```

```
    "confidence": "moderate",
    "support": "a simple argument that follows from the multiplicative structure",
    "residual": "convexity in real cost curves weakens the linear version"},
   {"memo": "Memo 3", "claim": "the response stage is under-supplied in these jurisdictions",
    "confidence": "low",
    "support": "expert opinion inherited from one comparator's experience (N14)",
    "residual": "jurisdiction-specific evidence not in the corpus"},
]
for r in register:
    print(f'{r["memo"]} | {r["claim"][:58]:58s} | {r["confidence"]:8s} | {r["support"][:44]}')

output_state["calibration_register"] = register
print("\nentries:", len(register), "| low-confidence entries:", sum(1 for r in register if r["confidence"] == "low"))
```

Three of six entries are low-confidence, and they are precisely the entries on which the recommendations turn. That distribution is the honest output of the drill. A memo that reported all six as moderate would be less transparent, not more decisive.

Closing note on what this drill cannot establish, stated so that it can be repeated verbatim in an interview. These are invented opportunities; nothing here is a reconstruction of a Coefficient Giving decision or evidence about how any real grant was evaluated. Real recommendations depend on private evidence, on conversations with applicants and third parties, and on portfolio context that a document cannot supply, and the corpus itself says so: P04 describes grant-makers asking early what evidence would change their mind and seeking it out, P04 and P04's colleagues describe investigations that run on calls and reference checks rather than on written proposals alone, and D04 says applications are evaluated holistically with no fixed number of grants intended. A drill can teach the shape of the reasoning. It cannot supply the evidence, and it should not pretend to.

**Final rehearsal instruction. For each memo, say the recommendation in one sentence, the decisive consideration in one sentence, the confidence and its support in one sentence, and the reversal condition in one sentence. If any of the four needs a second sentence, the memo is not yet finished.**

# Draft calibrated answers for comparing opposing arguments and expected value.

## Question

Draft a set of rehearsable answer scripts for the interview questions that test the role's stated evaluation criteria: comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and communicating concisely with calibrated confidence. Each script must state its assumptions, its confidence and its limitations, and must expose rather than conceal the premises it depends on.

Deliverable: a skeleton for any answer, three worked scripts for competing expert claims, two scripts for expected value with explicit unit discipline, two scripts for decisive considerations and proportionate diligence, and an adversarial set that defends the opposite conclusion at equal strength. Hard constraint: nothing may be invented. No application, credentials, employers, grants made or work history is asserted or implied; every simulation, threshold and parameter that is not stated in the corpus is labelled as invented and synthetic.

## Background

The role brief supplies the criteria directly: evaluation requires comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and judging people and organizations, with communication that is concise, calibrated and transparent about reasoning. The corpus supplies the raw material for scripts that satisfy those criteria without pretending to be fieldwork.

D01 (Reasoning Transparency) supplies the confidence vocabulary and the central warning that words such as *plausible* are interpreted differently by different readers, and it supplies a list of kinds of support ranging from a careful examination of studies you are qualified to assess down to a vague impression or a general intuition. It also recommends stating what could change the author's mind about the key takeaways. Those two elements become the calibration rails for every script below.

P03 is the corpus's model of a high-stakes quantitative argument: an explicit two-assumption model of frontier growth and lagged adoption, a stated central calibration, a headline welfare number in Coefficient Giving's own unit system, and a coda that separates AI-skeptic objections from AI-believer objections and responds to each. It also flags that its own scenario rests on huge ifs. P08 supplies the opposite-findings case: two funded groups produced results pointing in opposite directions on AI uplift to biological misuse, and the funder's response was to support both and let the ideas compete.

P12 supplies the hardest calibration problem for an evaluator's self-description:

everyone thinks they have good taste, there is no One True Taste, and taste may be better judged by whether it aligns with your values than by a universal ordering. L02 supplies the move structure and the observation that an evidential argument changes log-odds by an amount measured in bits, which gives a way to test whether a claim is even capable of changing a decision.

Standing caveats carry through every script. Institutional voice, individual-staff voice, third-party comparator voice and external primary material must stay separated; a per-staff giving suggestion does not represent the institutional view, an individual's informal timeline opinion is not institutional policy, and a comparator that operates the model it advocates has a stake in the answer. A dated number is a dated number: P08's 10% by 2036 estimate is a 2016 institutional anchor, not a present forecast.

## Sub-Question 1: The answer skeleton: one question, one crux, three numbers, one falsifier

Produce a reusable oral-answer skeleton that lets a listener reconstruct the question under discussion, the two opposing positions at their strongest, the crux, at most three quantitative anchors with units and provenance, my confidence and the kind of support behind it, and the observation that would move me. Then demonstrate it on a live timeline question, first at ninety seconds and then compressed to twenty, and show what the compression removed and why the listener can still supply it.

The timeline question is chosen because the corpus contains a deliberate voice conflict on it: two institutional sources reporting shortened timelines and treating fast progress as the default path, one institutional source whose only published probability is a decade old and whose own text records internal disagreement, one individual author who conditions everything on insiders being roughly right, and one informal individual opinion that timelines might be medium or broad.

### Step 1

The skeleton has six slots, and the discipline is that a slot may be omitted but never faked. If a slot has no honest content, say so in one clause and move on; an empty falsifier is itself information about how weak the claim is.

1. Restate the question in the asker's terms and name the decision it feeds. This is the first of L02's six moves and it is what D01 means by opening with the key takeaway.
2. State both positions at their strongest, in the source's own hedging, with the locus attached (institutional, individual staff, comparator, external primary).

3. Name the crux and say why it is a crux rather than a topic: something must change sign, or nothing has been decided by settling it.
4. Give at most three numbers, each with a unit and a provenance tag, and say explicitly which of them is my own arithmetic on someone else's parameter.
5. State confidence using the source's own wording where the source hedged, the source's own probability where the source gave one, and my own encoding only where I flag it as mine.
6. Name the falsifier: the class of observation that would move me, and roughly how much movement it would take. This is D01's 'what could change your mind' made operational.

Five calibration rules keep the skeleton honest.

- Never launder a hedge. If the source says may or could, the script says may or could. If the source gave only a directional update, the script reports a directional update.
- Never supply a number where the source supplied a word, and never cite a number without its unit. The register discipline from the previous phase treats both as failures, not as stylistic choices.
- Never encode a probability without naming the reference class. A probability of a world-state, a probability of a grant succeeding and a probability of an actor attempting harm are three different objects and are not commensurable.
- State the unit system before any comparative number: grant dollars, present-value income, welfare units, lives-equivalent. The same object appears in two unit systems in P03 and the two must not be compared to each other.
- Say what I am not claiming. An answer that lists its non-claims is self-limiting without being evasive, and it pre-empts the interviewer supplying an over-reading.

The skeleton earns its length only if its claims can supply enough evidential weight to move a close call. L02 says an evidential argument changes log-odds by an amount measured in bits, so the natural test of a drafted answer is whether any of its rows can supply more than a fraction of a bit.

$$\text{bits}(p_0 \to p_1) = \log_2 \frac{p_1}{1-p_1} - \log_2 \frac{p_0}{1-p_0}$$

The routine below encodes this and applies it to five candidate claims drawn from the timeline conflict, including one claim that has no numeric encoding at all and therefore cannot be counted as evidence weight. The bar is set at about

one third of a bit, roughly a likelihood ratio of 1.25, which is about the smallest update that can flip a genuinely close call.

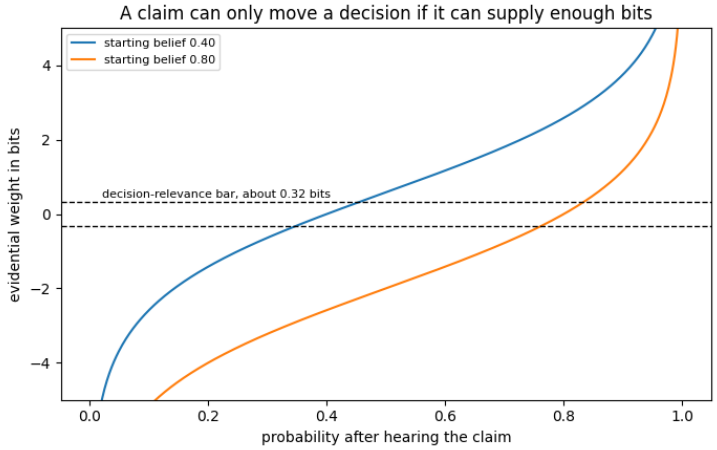


Figure 53: Graphical analysis for Step 1.

**Step 2**

Now apply the skeleton to a live question. The prompt is hypothetical and is marked as such: *Coefficient Giving has shortened its grant investigations and reports that its own staff updated toward shorter timelines. Does that change how you would spend a marginal grantmaker hour, and how confident are you in the underlying claim?*

Ninety-second version. The institution has two published positions that point the same way and one that does not. P01 and P06 both report shortened timelines, with P06 saying that continued fast progress increasingly looks like the default path; P08, which is an older institutional document, records a 2016 estimate of at least a ten percent chance of transformative AI by 2036 and records real internal disagreement. P03, an individual staff author, explicitly conditions its whole growth model on the AI-believer view being roughly right and flags at the end that these are huge ifs. An informal individual opinion in the corpus says timelines might be medium, or broad. My crux is not which timeline is true; it is whether a marginal grantmaker hour should go to work that pays off in both regimes or to work that only pays off in one. If the latter, the timeline question decides the hour. If the former, it does not. The decision-relevant evidence is therefore about regime-robustness, not about timelines. I would hold the two institutional registers separately rather than averaging them, and I would say plainly that I cannot convert P01's and P06's directional updates into a probability.

Twenty-second version. The institution is divided in the way the sources are divided: two recent institutional updates toward shorter timelines, one dated institutional anchor and an individual forecast that hedges heavily. The crux is whether a marginal hour should buy regime-robust work or a single-regime bet. I cannot price the timeline, so I would buy robustness and name the falsifier: capability measurement failing to extend would lower the weight I place on the short-timeline position.

Table 131: What survives compression, and what the listener can be trusted to supply. The crux and the falsifier are never cut; the numerics can be cut but their provenance and their date stay in the register.

| **Slot** | **90-second version** | **20-second version** |
|---|---|---|
| Question for the asker | How to spend a marginal grantmaker hour | kept |
| Positions A and B | Both institutional registers named with sources | compressed to one clause |
| Crux | Regime-robustness, not the timeline itself | kept, this is the load-bearing line |
| Numbers | Ten percent by 2036, P08, dated | dropped, but the date is not dropped from the register |
| Confidence and kind of support | Two directional moves, one dated probability, one hedged individual forecast | reduced to no priceable probability |
| Falsifier | Capability measurement failing to extend | kept, a script without a falsifier is not calibrated |

Finally, lint the script mechanically. The routine below checks that every slot is filled, that every number carries a unit and a provenance tag, that my own arithmetic is labelled as mine with a stated derivation, and that any confidence word is tied to the source hedge it came from. It is run on a deliberately flawed draft and then on a repaired version, so the failure modes are visible rather than hypothetical.

## Sub-Question 2: Scripts for competing expert claims, with locus preserved

Draft three answer scripts for questions of the form *two credible parties disagree; what do you think and why?* The three disagreements are the timeline and

urgency conflict inside and around Coefficient Giving, the opposite findings on AI uplift to biological misuse, and the dispute between centralised large-scale grantmaking and dispersed regranting. Each script must name the locus of each voice, state the crux, give the decisive consideration, and say what would move me.

**Step 1**

The three scripts share a shape and differ in their locus discipline. The comparison table below is the answer to the meta-question, *how do you handle disagreement?* because it shows that the handling is not one rule but a rule conditioned on who is speaking.

Table 132: The meta-answer: the handling rule is conditioned on locus, not on which claim sounds better.

| **Disagreement** | **Position A, at its strongest** | **Position B, at its strongest** | **Crux** | **Decisive consideration** |
|---|---|---|---|---|
| Timelines and urgency. Locus: P01 and P06 institutional; P08 institutional but dated; P03 individual staff; S11 individual informal. | Capability progress has outrun expectations; fast progress is the default path; the ecosystem should be seeded now because organisations need calendar time. | The published institutional probability is a decade old; internal disagreement is recorded in the same document; the quantitative author conditions on insiders being right and flags huge ifs. | Whether a marginal hour buys regime-robust work or a single-regime bet. | Calendar time is not recoverable: organisations need time to hire and plan, so today's dollar is worth more than a dollar in two years if the opportunity is real. |

| Disagreement | Position A, at its strongest | Position B, at its strongest | Crux | Decisive considera-tion |
|---|---|---|---|---|
| AI uplift to biological misuse. Locus: external primary, both groups funded by the institution. | Language models can help people with little or no lab training develop new pathogens, so the capability margin matters. | Current models do not substantially increase the risk of a biological attack once the full acquisition pathway is modelled. | Actor capability, access and the counter-factual baseline are being held at different values in the two studies. | The institution plans for the possibility that models approach biological-weapon risk thresholds, and funds both sides so the best idea wins. |
| Centralised versus dispersed grantmak-ing. Locus: P06 institutional versus P11 and P12 third-party comparators with a stake in the model they advocate. | A small army of specialised full-time grantmakers, light staffing ratios, and moving many dollars and many grants per staff member. | Many people's first hundred thousand dollars of grants will beat a professional grantmaker's last hundred thousand; regrantors are cheap to scale and better at early-stage dealflow. | Whether taste and dealflow scale better dispersed than centralised. | Marginal returns: the case for dispersal is strongest at the smallest grant sizes and weakest where diligence cost per dollar is genuinely high. |

Two rules of delivery apply to all three. First, concede the strongest version of the opposing case before defending my own, because the brief asks for comparing opposing arguments rather than defeating one. Second, never resolve a disagreement that the evidence does not resolve: the previous phase's finding is that holding C1-1 and C1-2 open while taking a position on C2-1 is the correct posture, because the first two are genuine cruxes and the third is a comparative question with a live empirical answer.

### Step 2

The biorisk script needs one extra move, because a naive treatment turns a conditional disagreement into a contradiction. Write the risk as a product and ask which factor each study is moving.

$$\text{risk} = P(\text{acquisition}) \times P(\text{success} \mid \text{acquisition}) \times \text{consequence}$$

*The uplift finding is about the middle factor, evaluated on an actor who has already committed to attempting harm and who has little or no lab training. The null finding is about the whole product, evaluated on a prospective attacker inside a realistic acquisition pathway. Both can be true simultaneously, and the institution's own position supports funding both. The illustrative arithmetic below is synthetic and invented purely to make the reconciliation visible in numbers; none of it is an estimate from the corpus and none of it should be quoted as one.*

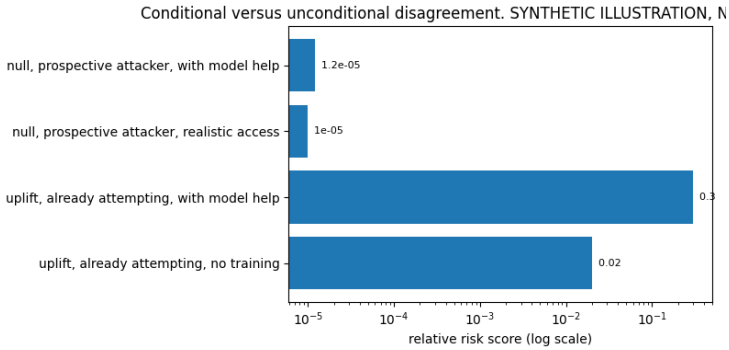


Figure 54: Graphical analysis for Step 2.

## Sub-Question 3: Scripts for expected value: unit discipline before arithmetic

Draft expected-value scripts for the questions the brief names: reasoning quantitatively and probabilistically about expected value, and reasoning under uncertainty generally. The scripts must show the model, show the sensitivity, state the reference class, and refuse to compare numbers that live in different unit systems.

### Step 1

The first expected-value script answers a portfolio question: at a given grant size, how many grants does a hits-based program need before it can expect a

hit, and what does that imply about the grant-size band it should operate in? The corpus supplies the shape of the answer. P11 states the hits-based logic explicitly and gives a worked retrospective: a single small early grant grew into a stake worth many multiples of the whole regranting budget. P10 states the same logic prospectively, that early funding tolerates failure and that payoffs are power-law distributed.

$$n = \left\lceil \frac{\ln(1-q)}{\ln(1-p)} \right\rceil, \qquad \bar{g} = \frac{B}{n}, \qquad \mathbb{E}[V] = pR$$

The script's caution is that expected value is only as good as the reference class behind the hit probability. A hit probability estimated from one program's own retrospective is not a base rate for another program, and the closing constraint from the failure-analysis work applies here directly: a single success story demonstrates a mechanism but cannot estimate how often the mechanism fires.

The routine below computes the portfolio arithmetic for a two million dollar budget across hit probabilities from three in a thousand to three in a hundred, at three confidence targets, and reports the average grant implied in each case. The reading is that a program wanting a ninety percent chance of at least one hit at a one percent hit rate needs on the order of two hundred and thirty grants, which forces an average grant near nine thousand dollars: that is the arithmetic behind the small-grant niche, and it is the reason the grant-size band is not a free choice.

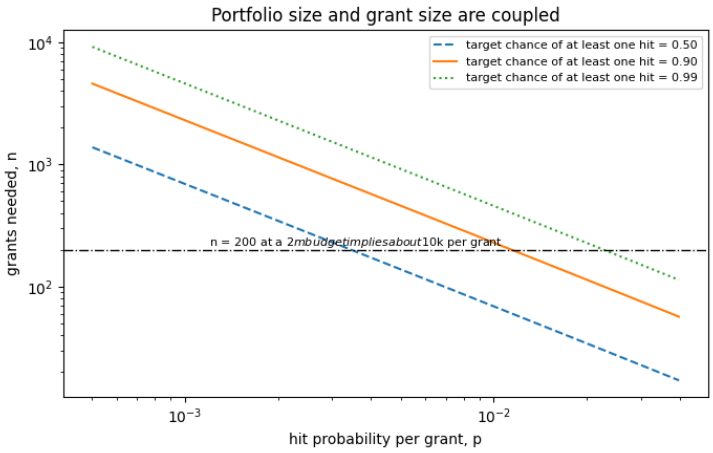


Figure 55: Graphical analysis for Step 1.

**Step 2**

The second expected-value script is about unit discipline, and it exists because the corpus contains two headline numbers for the same object in two different unit systems. P03 reports a present-value income figure of about forty-two point six trillion dollars per year of faster adoption, and separately reports the same comparison in the funder's welfare units at about three hundred and eighty-one trillion CG dollars per year, equivalent to about one hundred and fifteen million lives. The welfare methodology it cites is that a one percent income increase for one person for one year is worth five hundred CG dollars.

$$\text{CG\$ per life-equivalent} = \frac{6.1 \times 10^{15}}{1.86 \times 10^{9}} \approx 3.3 \times 10^{6}$$

The check below divides the total welfare gap by the total lives-equivalent, and divides the per-year welfare figure by the per-year lives figure, and compares the two. They agree to about one percent, which is the rounding resolution of the published headline numbers themselves. The non-trivial consequence for an interview answer is the interpretation: a life-equivalent in this accounting is roughly six and a half thousand person-year-percent units, so it is a welfare accounting unit and not a mortality statistic. Quoting it as deaths averted would be a category error, and that is exactly the kind of laundering the calibration rules forbid.

**Step 3**

The deliverable for this subquestion is a single spoken expected-value answer, assembled from a provenance table. The table is the answer's backbone: every number a listener hears is in it, with its unit, its source, and whether it is source-stated or my arithmetic.

Table 133: The provenance table is the answer. Reading it aloud with units and sources is the script; the prose merely connects the rows.

| Number | Unit | Source | Provenance and caveat |
|---|---|---|---|
| about 230 grants | count | my arithmetic | Derived from a chosen hit probability of one percent and a ninety percent target; the input probability is my assumption, not a measured base rate. |
| about 8,700 dollars | dollars per grant | my arithmetic | Two million dollar budget divided by 230. The budget is an illustration, not a program's actual budget. |
| up to 10,000 dollars | dollars per grant, ceiling | D03 and S06 | A live ceiling for one small-grant program, with average decisions inside a day and payment upfront by default. |
| below about 150,000 dollars | dollars per grant, typical band | P10 | Stated by a comparator fund describing its own comparative advantage, so it is a self-description. |
| 200,000 to 2,000,000 dollars per year | dollars per year, preferred first grant | D04 | The institution's stated preference, with projects below that scale ideally funded elsewhere. |

| Number | Unit | Source | Provenance and caveat |
|---|---|---|---|
| 500 CG dollars | welfare units per one percent income rise for one person for one year | P03 | The methodology the author uses to value income gains; it is an accounting convention, not a moral claim anyone has to accept. |
| about 3.3 million CG dollars | welfare units per life-equivalent | my arithmetic on P03 | Two independent checks on the source's own rounded headline numbers agree to about one percent. |

The spoken version, in four sentences: the grant-size band is not a free choice, because at a plausible hit probability a portfolio that wants a ninety percent chance of a hit needs hundreds of small grants, which forces an average grant around ten thousand dollars; that is why the smallest programs in the corpus operate at that scale and why they must decide fast, with a stated average decision inside a day and payment upfront by default; the large institution's stated preference for first grants in the two hundred thousand to two million per year range is consistent with the same arithmetic once its evaluation cost per dollar is higher; and I would not compare the small program's average grant to the large program's first-grant floor unless I also compared their diligence cost per dollar, because the two are jointly determined.

## Sub-Question 4: Scripts for decisive considerations and proportionate due diligence

Draft two scripts for the brief's remaining criteria: identifying decisive considerations, and choosing proportionate due diligence. The first must convert a crux into a numerical threshold and state how much evidence it would take to cross it. The second must show that proportionality is arithmetic rather than taste, and must reproduce the grant-size banding the corpus already displays without being told to.

### Step 1

A decisive consideration is not a topic, an uncertainty or an important-sounding variable. It is a premise whose value changes the sign of the difference in value between acting and not acting.

$$\Delta V(p) = V_{\text{act}}(p) - V_{\text{omit}}(p), \qquad \Delta V(p^*) = 0$$

The illustration below is hypothetical and labelled as such. Two strategies are compared as a function of the probability that transformative AI arrives inside the planning horizon. Strategy A is smaller but robust in both regimes; strategy B is larger if takeoff is slow and much smaller if takeoff is fast. The crux is the crossing probability, and the answer is not the crossing probability itself but the distance between my current belief and that crossing, expressed in bits using the routine from the skeleton step.

The script sentence this produces is: the recommendation is stable as long as the median probability stays below the crossing value; my current median implies a gap of roughly one and a half bits, so I would need a class of observation capable of supplying more than that before I would change the recommendation. That sentence does three things at once: it names the crux, quantifies the distance to the flip, and states what evidence would be required without pretending to know whether such evidence exists.

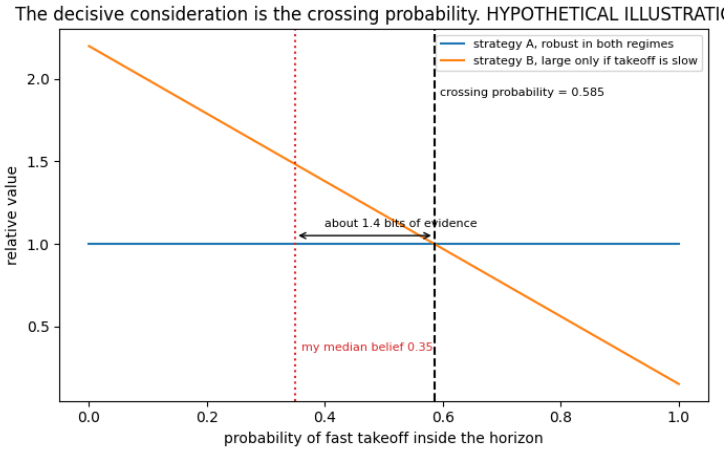


Figure 56: Graphical analysis for Step 1.

### Step 2

Proportionate due diligence is the same arithmetic applied to investigation rather than to funding. An investigation is worth its cost only if it changes

the probability of a bad outcome enough to matter at the downside being protected against.

$$\mathbb{E}[\mathrm{VOI}] = \left(p^{0}_{\mathrm{bad}} - p^{1}_{\mathrm{bad}}\right) L - c, \qquad L^{\text{break-even}} = \frac{c}{p^{0}_{\mathrm{bad}} - p^{1}_{\mathrm{bad}}}$$

The crux of this script is that the break-even downside scales linearly with investigation cost and inversely with the information the investigation actually supplies. The routine below tabulates break-even downsides for three investigation costs and three sizes of probability reduction. The important output is not any single number but the shape: at a small investigation cost the break-even downside sits in the tens of thousands of dollars, and at a large investigation cost it sits in the hundreds of thousands.

This shape reproduces the corpus's own banding without being fitted to it. A program with a ten thousand dollar grant ceiling and same-week decisions is operating where a full investigation could not pay for itself; the large institution's stated preference for first grants in the two hundred thousand to two million per year range is operating where it can; and the operational descriptions in the corpus match, with one grantmaker holding six to twelve investigations at various levels of focus and another describing a habit of asking early what piece of evidence would most likely change his mind, then seeking that specific evidence in a short call rather than running a full review.

The caution I would state aloud: the break-even loss is a lower bound on the case for investigating, not a licence to under-investigate. Two adjustments push it up: the same hours could have been spent on a larger opportunity, which is the institution's own stated tension between steady support for existing grantees and proactive seeding of new organisations; and some failure modes are not captured by a probability of funding a bad grant at all, because they are reputational or ecosystem-level, which is why the institution describes itself as unusually careful about reinforcing best practices across the field.

## Sub-Question 5: Adversarial calibration: reversing each script and stating the limits

For each of the three competing-claims scripts, write the opposite conclusion at equal strength and compute what it would take to reverse the recommendation. Then state the limits of the whole answer set: which parts are source-supported, which parts are my arithmetic on source-anchored parameters, which parts are invented illustrations, and what I would have to be told before these scripts should be discarded.

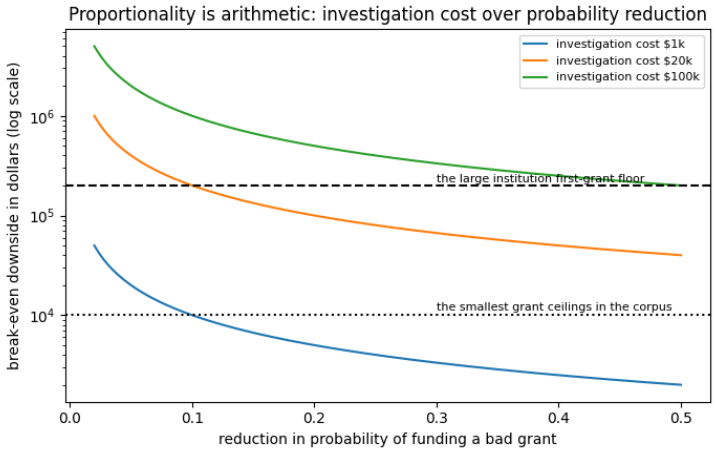


Figure 57: Graphical analysis for Step 2.

**Step 1**

A confidence stated in an answer creates an obligation to say what would overturn it. The standard form is a likelihood ratio: the amount of contrary evidence needed to reach the point of indifference, and the amount needed to put the opposite conclusion at the same strength the original had.

$$\text{LLR to indifference} = \frac{p}{1-p}, \qquad \text{LLR to reverse at equal strength} = \left(\frac{p}{1-p}\right)^2$$

The routine below applies this to the three defended positions and names a class of evidence that could plausibly supply the required weight. The pairing of a required weight with a named evidence class is the part that converts a slogan about being open-minded into a falsifiable commitment.

The three reversals, stated so they can be defended: if capability measurement plainly stopped extending for a sustained period, the short-timeline position loses its main support and the case for a broad, hedge-across-timelines portfolio weakens; if the uplift finding failed to replicate under realistic access controls and the null finding replicated on the same conditions, the biorisk capability margin becomes a lower priority; and if a costed, counterfactually-adjusted comparison of regranted and centrally allocated dollars at the same stage showed no advantage to dispersal, the regranting case reduces to a preference about values rather than a claim about marginal returns. P12 anticipates the third of these when it says there is no One True Taste and that the question may be whether a given taste aligns with your values rather than whether it is better.

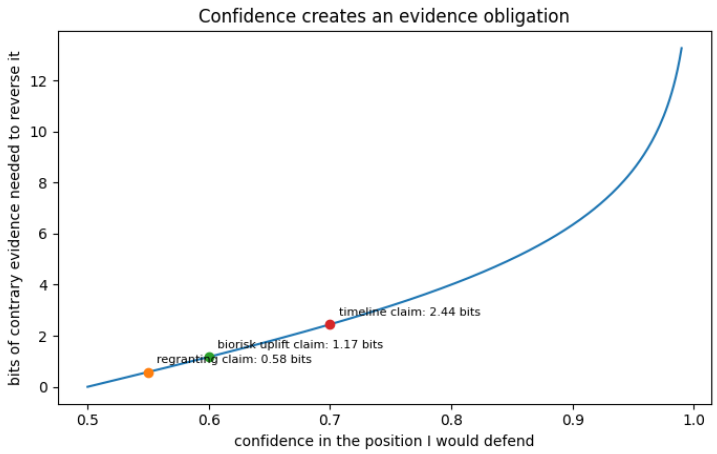


Figure 58: Graphical analysis for Step 1.

**Step 2**

Limits of the answer set, stated in the order the corpus would state them.

- Source-supported: the evaluation criteria come from the role brief; the confidence vocabulary, the kinds of support and the warning about words like plausible come from the transparency document; the biorisk opposite findings and the decision to fund both come from the institutional AI-safety essay; the grant-size examples come from the small-grant program page, the comparator fund launch post and the governance request for proposals; the unit methodology and the takeoff model come from the catch-up-growth post.
- *My arithmetic on source-anchored parameters: the portfolio size and average grant, the welfare unit consistency check, the crossing probability, the bits-to-threshold, the break-even downside and the reversal weights. These can be wrong if I have mis-set an input, and the inputs include numbers I chose rather than read.*
- *Invented illustrations: the two-strategy payoffs and the belief distribution in the decisive-consideration step, and the four biorisk scenario probabilities. Both are labelled in the code and in the prose and neither may be quoted as an estimate.*
- Voice risks I have tried to close: reading an individual staff member's growth model as the institution's position, reading a per-staff giving suggestion as institutional, treating a comparator that operates the model it advocates as neutral, and treating a dated institutional probability as a present forecast.

- *What is not claimed anywhere in this pack: no application has been made, no credentials, employer, grant or work history is asserted or implied, and no example depends on a track record I have not been given.*

What would make me discard the skeleton itself. If it turned out that listeners weight a named confidence word more than a stated one, the confidence slot should default to numbers with an explicit reference class. If it turned out that interviewers treat a stated falsifier as indecision rather than calibration, then the falsifier should be compressed into a clause inside the crux slot rather than given its own slot. If it turned out that the crux slot is usually empty in practice, then the skeleton should be reordered to put the decisive consideration before the two positions, because in that case the positions exist to justify the consideration rather than the reverse.

The final self-assessment question I would put to myself before using any of this: which sentence in the answer am I most confident about, and is that confidence earned by source support or by the fact that I wrote it? The transparency document's own warning applies reflexively here, because the person who has just produced a skeleton is exactly the person most likely to overstate how much the skeleton settles.

# Practice concise communication and reasoning transparency.

## Question

This step converts the preceding synthesis into interview performance: three 60-120 second spoken answers and two short written memos. Each deliverable must open with linked key takeaways, state confidence and kinds of support, name cruxes, and end with a clear recommendation. The constraint is strict: no invented credentials, work history, experience or application; every hypothetical must be labelled.

**The non-trivial jump is not formatting. It is deciding which considerations are decisive, how much confidence to express, and which support kind you actually have, then cutting the answer until every remaining sentence helps the listener take the next inferential step.**

## Background

Role brief communication norms: the evaluator role requires concise, calibrated and transparent reasoning; evaluation requires comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, and judging people and organizations.

D01 recommends opening with a linked summary of key takeaways, indicating which considerations are most important, and stating confidence and kinds of support. It gives examples ranging from 'seems likely' to explicit probabilities and confidence intervals.

L02 frames writing as arranging affordances for the reader: meet the reader where they are, take them to the claim, show how you could be moved, show the evidence, apply the criterion, and state the update. Its warnings include unnecessary contrast, buried referents, abstraction that hides human agency, and variation that creates interpretive work without purpose.

D02 is a counterexample to over-reliance on AI writing: it says AI tools remain poor at serviceable prose, that output can be sales-y and slop-like, and that apparent speed can mask lack of progress. Use AI for synthesis and source gathering; use your own voice for final answers.

**The source index for this step is the role brief; D01, D02, D04, L01, L02; P04, P06, P08, P10-P17; N01-N06 and N14; S01-S11. Source-supported claims are attributed by ID. Proposed exercises are labelled as inferences.**

## Sub-Question 1: Build the Communication Contract: D01 plus L02 as a Checklist

Deliverable: a one-page checklist and timing template that every spoken answer and memo must satisfy. This is a proposed exercise.

### Step 1

Map D01 requirements to L02 moves and to the surface of an interview answer.

Table 134: Inference: this mapping is a proposed synthesis of D01 and L02, not a source statement.

| **D01 requirement** | **L02 move** | **Answer surface** | **Failure if absent** |
|---|---|---|---|
| Open with linked summary | Move 1: meet reader where they are | First 10 seconds: takeaway and signpost | Listener cannot restate the claim. |
| Indicate most important considerations | Move 3: show how you could be moved | Names one or two decisive considerations and the crux | Crux becomes a topic, not a test. |
| Indicate confidence and support | Moves 4-5: evidence and criterion | Gives probability or qualifier plus support kind | Numeric laundering. |
| Linked summary throughout | Move 2: take reader to claim | Signposts where the claim is developed | Vague assertion. |
| State update | Move 6: show where reader should now be | Ends with recommendation and what would change mind | No decision. |

### Step 2

Name the non-trivial failure modes and the test that catches each one.

Table 135: Proposed exercise: these tests are inferred from the sources, not printed in them.

| **Failure mode** | **L02 or D01 warning** | **Test** | **Repair** |
|---|---|---|---|
| Unnecessary contrast | L02: 'This is X, not Y' introduces Y without premise | Ask listener what alternative was in the question | Cut Y or establish the premise. |
| Buried referent | L02: keep the gun close to the shot | Underline pronouns; antecedent within one sentence | Repeat the noun. |
| Abstraction hides agency | L02: human primacy over agency | Name who decides, assumes, estimates | Make the person the grammatical subject. |
| Confidence laundering | D01: express degrees of confidence and support kinds | If 'likely', give source's own string or rough probability | Say 'no basis' if none. |
| Voice merging | Prior phase: attribution ladder | Label institutional, individual staff, third-party, external | Add locus. |
| AI prose slop | D02: AI is poor at serviceable prose and speed can mask no progress | Read aloud; if sales-y, rewrite | Use AI for synthesis, not final voice. |

**Step 3**

Set a 60-120 second timing template.

Table 136: Inference: the exact second counts are a rehearsal scaffold; D01 and L02 support the components.

| **Time** | **Purpose** | **Content** |
|---|---|---|
| 0-10 sec | Linked takeaway | State recommendation and the decisive consideration. |

| Time | Purpose | Content |
|---|---|---|
| 10-25 sec | Meet reader | Name shared criterion: theory of change, track record, cost-effectiveness. |
| 25-45 sec | Claim and crux | Give your claim and the premise that changes the decision. |
| 45-75 sec | Evidence and support | Name source IDs and support kind. |
| 75-100 sec | Apply criterion | State confidence update and counterevidence. |
| 100-120 sec | Recommendation | End with action and what would change your mind. |

## Sub-Question 2: Three 60-120 Second Spoken Answers

Deliverable: three spoken answers. Each is a rehearsal script with linked takeaway, confidence, support kinds, crux and recommendation. None asserts personal credentials.

### Step 1

**Answer A: evaluating a proposal in an area where experts disagree.**

Rehearsal script. No personal claim; this is a practice construction using source-supported reasoning.

Question: How would you evaluate a proposal in an area where experts disagree?

Takeaway: I would not average the experts. I would identify the crux and buy only the evidence that moves it. Confidence: moderate on the method; low on any specific net-risk estimate. Support: D01 recommends indicating confidence and support; P08 records that Open Philanthropy funded opposing papers and awarded a contest to be moved; D04 lists theory of change, track record, strategic judgment, project risks and cost-effectiveness as assessment criteria; N14 shows that political feasibility can fail even when the technical case is sound.

Crux: whether the proposal's theory of change changes sign under the disagreement. If the disagreement is about timelines, the decisive question is whether the intervention is robust across short and long timelines.

Evidence I would seek: a written threat model, a named expert whose disagreement would change my mind, and a 90-day observable that distinguishes the branches.

Recommendation: if the proposal is robust and the crux is observable inside the grant period, recommend a small first grant with a kill criterion; if not, decline and refer to a funder whose mandate matches. What would change my mind: an expert I trust showing the crux is already settled or that the theory of change collapses under one branch.

**Step 2**

**Answer B: weighing centralized large grantmaking against dispersed regranting.**

Question: How do you weigh centralized large grantmaking against dispersed regranting?

Takeaway: the portfolio needs both; the question is which decision each mechanism is better at. Claim: centralized grantmakers have comparative advantage at large, inframarginal, founder-recruiting grants; dispersed regrantors have comparative advantage at small, fast, early-stage discovery. Confidence: high on the structural distinction; moderate on the optimal mix.

Support: P06 defends Coefficient Giving's light staffing ratio and argues that moving grantmaking staff from payroll to a grantee's is not necessarily a big efficiency win, while agreeing regranting can be very impactful. P11 and P12 argue that many people's first $100k of grants will beat a professional grantmaker's last $100k and that regranting scales cheaply and seeds opportunities early.

Crux: whether taste and dealflow scale better dispersed than centralised. The observable is comparative risk-adjusted return by grant size and discovery source.

Evidence: P11 cites a $143k regrant into Timaeus that grew into a stake later associated with a $160m grant, but that is one hit, not a rate. I would want follow-on funding rates and failure rates for regrantor portfolios.

Recommendation: use centralized capacity for large scale-ups and proactive seeding; use regrantors for small exploratory grants with pre-committed top-ups; measure discovery source and follow-on funding. What would change my mind: evidence that dispersed regrantors systematically underperform after selection effects are accounted for.

**Step 3**

**Answer C: judging a founder or team when the main risk is execution capacity.**

Question: How do you judge a founder or team when the main risk is execution capacity?

Takeaway: I would test for the specific bottleneck, not general intelligence. Claim: the decisive consideration is whether the team has a credible plan to

relieve the binding constraint. Confidence: moderate; references can be noisy and the base rate information is thin.

Support: P04 shows grantmakers asking what piece of evidence would change their mind; Trevor Levin asks early and often; Chris Bakerlee checks track record, theory of change and downside risks; P12 decomposes grant programs into money, taste, dealflow, hustle and trust.

Crux: whether the founder is spread too thin or has the agency to hire and delegate.

Evidence: ask for a workplan with named owners, a 90-day milestone, and a reference who has seen them execute under load; if the right expert would change the judgment, take the 15-minute call that P04 describes.

Recommendation: if the bottleneck is founder attention, fund a smaller pilot with a hiring milestone or bridge support to recruit a number two; if the bottleneck is idea quality, decline. What would change my mind: a reference showing the founder has previously delegated and scaled.

## Sub-Question 3: Two Short Written Memos

Deliverable: two memos. Each has linked key takeaways, confidence statements, kinds of support, a crux and a recommendation.

### Step 1

**Memo 1: Funder-fit and recommendation for two labelled hypothetical opportunities.**

Linked key takeaways. Takeaway 1 links to the criterion section: funder fit is not a quality judgment; it is a mapping from an opportunity to a mandate, capital band, evaluation cost and risk posture. Takeaway 2 links to the recommendation: the two hypotheticals should not be forced through the same process. Takeaway 3 links to the confidence note: the fit logic is stronger than any quality inference.

Hypotheticals. Opportunity A is a two-person team with a working prototype requesting $60k for a six-month AI governance research pilot, no institutional home, one BlueDot alum. Opportunity B is an established 501(c)(3) with a three-year track record requesting $1.5m per year over two years to scale a safety evaluation program. Both are labelled hypotheticals from the preceding phase.

Recommendation: refer Opportunity A to the Transformative AI Fund, Manifund regranting, or BlueDot Rapid Grants if staged to $10k or less and community-eligible; do not route it to Coefficient Giving as a first ask unless it is pooled into a larger institutional proposal. Recommend Opportunity B to Coefficient Giving or another large funder if it clears the $200k-$2m/yr preference in D04.

Confidence: high on fit logic; low on actual quality because the hypotheticals have no source evidence. Support: P10 says the Transformative AI Fund primarily funds early-stage projects with grants typically below $150k and rarely above $300k; D03 says BlueDot Rapid Grants run from $50 to $10,000; D04 says Coefficient Giving typically prefers a first grant of $200k-$2m/yr over one to two years; P11 and P12 describe regrantor pots around $100k-$200k.

Crux: whether Opportunity A can be staged into a small funder's band without destroying the pilot, or matched by a regrantor. The decisive observable is a costed budget with a single primary deliverable and a named cost driver.

Evidence: the prior phase's arithmetic places BlueDot, LTFF, ACX and Manifund pots in different bands; the $60k request exceeds BlueDot's ceiling but fits a regrantor or the Transformative AI Fund. What would change my mind: a co-funder committing the balance, or a costed plan showing the pilot can be split into a $10k first milestone.

**Step 2**

**Memo 2: Conditional disagreement on AI biorisk uplift.**

Linked takeaway: the apparent contradiction between AI uplift findings and RAND's no-substantial-risk finding is a conditional disagreement, not a symmetric dispute.

Confidence: moderate-high on the reconciliation; low on the net effect. Support: P08 says both papers were funded by Open Philanthropy and point in opposite directions regarding AI's impact on biorisk; it describes a paper finding that LLMs could help people with little or no lab training develop new pathogens, alongside a RAND paper finding current language models did not substantially increase the risk of a biological attack.

Claim: the two findings model different actors, capabilities and counterfactual baselines. The uplift finding is about novice-assisted knowledge elicitation; the RAND finding is about a realistic acquisition pathway and end-to-end attack risk.

Crux: whether the operative risk is novice uplift or a capable actor using AI to overcome a specific acquisition bottleneck. That determines which safeguard matters: knowledge controls, acquisition controls, or both.

Evidence: read both papers and extract actor, capability, baseline and outcome. If the uplift paper uses a baseline of no lab training and the RAND paper uses a baseline of existing access, they are compatible.

Recommendation: do not average the findings. For a grant investigation, require proposals to name which branch they address and to specify the counterfactual. Fund work that resolves the crux, such as a threat model with a defined attacker profile. What would change my mind: a study showing uplift persists for a capable actor with lab access and an acquisition pathway.

## Sub-Question 4: Rehearsal, Feedback and AI-Writing Guardrails

Deliverable: a rehearsal protocol that stress-tests the spoken answers and memos without relying on jargon or unstated inference steps.

### Step 1

After each spoken answer, ask the listener to restate four things: your claim, your crux, your confidence and support, and your recommendation. Do not correct them immediately. Their reconstruction identifies the missing inferential step.

Table 137: Proposed exercise: this feedback loop is inferred from L02's reader-affordance model.

| **Listener restatement** | **What it diagnoses** | **Repair** |
|---|---|---|
| Claim | Whether the linked takeaway survived | Move takeaway earlier. |
| Crux | Whether the decisive consideration was audible | Name the observable. |
| Confidence and support | Whether you sounded calibrated | Give source's own string or say no basis. |
| Recommendation | Whether the update was clear | End with action and what would change mind. |

### Step 2

Record the answer. Transcribe it. Cut the word count by half without deleting a necessary connection. Then read the short version aloud. If the listener can still restate claim, crux, confidence and recommendation, the cut was safe. If not, restore the sentence that carried the connection.

- Cut throat-clearing and restatements of the question.
- Cut unnecessary contrast when the alternative was never in play.
- Cut adjectives that do not change the decision.
- Cut hedges that do not carry a probability or support kind.

### Step 3

D02 warns that AI remains poor at serviceable prose and that apparent speed can mask lack of progress. Use AI for synthesis, source gathering and BOTEC generation, but use your own writing for final answers.

Table 138: Suggested exercise: deliver the three answers aloud twice, once at 120 seconds and once at 60 seconds, and have a listener complete the restatement table. Do not memorise the memos; use them to practise the structure under time pressure.

| **Criterion** | **0** | **1** | **2** |
|---|---|---|---|
| Linked takeaway | Absent | Present but vague | First 10 seconds and listener can restate |
| Decisive considerations | List | One named | One named and crux test observable |
| Confidence and support | No qualifiers | Qualifiers | Qualifier plus support kind |
| Clear recommendation | No action | Action | Action plus what would change mind |
| Voice discipline | Merged | Partly labelled | Institutional, individual and third-party labelled |

# Rehearse institutional-disagreement and uncertainty questions.

## Question

You will be asked, in some form, whether a contested claim is true, what Coefficienct Giving thinks about it, how confident you are, and what would change your mind. These questions are graded on structure more than on conclusions: can you hold a disagreement open without resolving it by averaging, can you state the opponent's best case, can you name the premise that decides between the two, and can you say what evidence would move you.

Two hard constraints apply throughout. First, no invented credentials, history or experience: every position below is sourced, and every hypothetical is labelled as one. Second, organisations and individuals must not be merged into a single voice. A CEO letter, a guest post by a grantmaker, a third-party fund's launch post and an unfunded grant application are four different kinds of evidence, and treating them as one is the most common failure mode this set is built to catch.

The rehearsal below covers seven drill prompts plus one capstone circuit. Each drill gives you a source-anchored position, a charitable steelman, the crux expressed as a testable premise, and the observable that would move you.

## Background

What the corpus supplies for this exercise. It contains real disagreements, not constructed ones: an informal disagreement list (S11); a direct comparison of centralised and dispersed grantmaking with an explicit trust and principal-agent discussion (P12, P11); a post built around its own AI-skeptic and AI-believer objections (P03); a CEO letter that concedes the accuracy of people with shorter timelines than his own while warning that extrapolative expectations can drive bubbles (P06); two funds with deliberately different grant-size bands and speeds (P10, P11); an institutional commitment to funding experts who disagree with each other, illustrated by two funded papers that point in opposite directions on AI and biorisk (P08); open questions published by a grantmaking programme (P15, P17, P05); and partial maps with stated omissions (N06, N14).

What the corpus does not supply. It does not contain a distribution over staff timelines, the primary papers behind the biorisk summary, outcome data on declined grants, or any measure of how often the mechanisms described in the grantmaking-structure debates actually fire. Every drill therefore ends with a residual: the thing you would have to ask for rather than assert.

The one piece of formalism used throughout: a crux is the premise whose truth flips the recommendation, not merely a topic on which people differ. Write it as a sign change.

$$\Delta V(p) = V_{\text{act}}(p) - V_{\text{omit}}(p) \quad \text{flips sign at } p^*$$

If no reasonable value of p produces a sign change, you do not have a crux; you have a topic. Say so, and move to the next premise.

## Sub-Question 1: Drill 0 — the rehearsal form and the attribution discipline

Before practising content, fix the form. Every answer in this set is built from six moves. Locate (who is speaking, at what attribution level). Position (the claim in one sentence, with scope). Steelman (the strongest version of the strongest opposing source, in that source's own terms). Crux (the premise that flips the recommendation, with a criterion). Mover (the observable that would change your answer, and how much change it licenses). Residual (what the corpus cannot settle, and what you would ask for).

Three traps recur and are worth naming explicitly, because naming them in an interview is itself a signal of calibration: merging voices (treating an individual's post as institutional policy), confusing dated with current (a 2016 estimate inside a pre-rebrand post is an anchor, not a forecast), and treating a self-interested comparator as neutral (funds that advocate the fund form, or regranters who advocate regranting, are describing a model they operate).

### Step 1

The attribution ladder has four rungs. Each rung licenses a different sentence, and the sentence is the thing being graded.

Table 139: One observed claim can sit on several rungs at once: a signed institutional post can carry a first-person update. Record both.

| Rung | What it licenses you to say | Corpus examples |
|---|---|---|
| Institutional | This is the organisation's stated stance, subject to its date and to whether a newer published position supersedes it. | Programme post on scaling AI and biosecurity (P01); the CEO letter (P06); the AI safety and security approach, published under the former name (P08); fund pages (P05, D04); the reasoning-transparency note (D01). |

| Rung | What it licenses you to say | Corpus examples |
|---|---|---|
| Individual staff | This is one person's model, suggestion or opinion. It is admissible as evidence of how someone in the field reasons; it is not admissible as the institution's view. | The catch-up-growth modelling post (P03); the per-staff donor suggestions that state they do not represent the institutional view (P07); the founder call (S09); the informal disagreement list (S11). |
| Third-party comparator | This is a live strategic alternative with its own incentives. Cite it as an alternative and flag the interest. | The Transformative AI Fund launch (P10); Manifund regranting (P11) and the money, taste, dealflow, hustle, trust post (P12); BlueDot rapid grants (P13, D03). |
| External primary | This is material authored elsewhere and cited here. Useful for the territory, not for the institution's positions. | The governance map with seventeen activities and stated omissions (N06); the policy-feasibility retrospective (N14); the competition-modelling grant (N02). |

Encode the discipline so it survives under interview pressure. The card below refuses to validate unless every cited source has a recorded locus, unless position and steelman cite different sources, and unless you have written down what would not move you.

```
REQUIRED = (
    "drill_id", "question", "position", "position_sources",
    "steelman", "steelman_sources", "source_loci",
    "crux", "crux_criterion", "mover", "mover_observable",
    "mover_direction", "not_a_mover", "residual", "status_tag",
)
```

### Step 2

Two further disciplines make the difference between a rehearsal and a recitation.

- Scope the position to five parameters before arguing for it: the actor modelled, the capability modelled, the outcome measured, the counterfactual baseline, and the decision the claim bears on. Most apparent disagreements in this corpus are differences along one of those five, not differences about the world.
- Make the criterion falsifiable, in the sense the writing-structure source uses for a methodology move: it must allow a meaningful possibility that you are wrong. Asking an opponent for an impossibly complete alternative keeps your position alive by construction, and an interviewer will notice.

```
def check_mover(mover, direction, threshold, would_not_move):
    """A mover is only a mover if it can come out both ways."""
```

#### Step 3

Finally, budget the answer. A disagreement question in an interview is usually scoped to about ninety seconds. Allocate words deliberately, with the steelman getting the largest share, because the steelman is the part that cannot be bluffed.

At a normal speaking rate of roughly 2.2 to 3.0 words per second, ninety seconds is about 200 to 270 words. Measured in that budget: position about 30 words, steelman about 60 to 80, crux about 40, mover about 40, residual about 30. Anything longer means one of the moves has been dropped.

## Sub-Question 2: Drill A — are short timelines the default path?

Prompt, in the form an interviewer might use it: your programme's own writing treats rapid progress as the default, and the granting has scaled to match. Is that right, and how would it change what you fund first?

The trap here is treating a question about the world as if it were a question about the institution. The two have different answers and different evidence.

#### Step 1

Locate and state. Do not average the sources into a single house view; they carry different loci and different dates.

Table 140: Five sources, four rungs, three dates. The disagreement is not symmetric: the long side here is represented mostly by conditional modelling and informal opinion, while the short side is represented by two institutional updates.

| **Source and locus** | **Claim, in its own framing** | **Date and status** |
|---|---|---|
| P01 — institutional post carrying a first-person update | Timelines have shortened; the people taking short timelines seriously look increasingly correct; most of the work is worth doing even under long timelines. | Current. |
| P06 — institutional, CEO letter | Short-timeline predictors have been more accurate than mine; capability measurements continue up and to the right; continued fast progress increasingly looks like the default rather than an aberration. | Current, and itself records a two-year sequence of being wrong in the same direction. |
| P08 — institutional, published under the former name | At least a 10 percent chance of transformative AI by 2036; substantial expert disagreement remains, and many staff now put higher odds on 2035. | **The estimate is from 2016. Dated anchor only; do not quote it as a current forecast.** |
| P03 — individual staff modelling | Mainstream economists put AI-induced excess growth at 0.1 to 1.5 percent per year; the whole scenario conditions on insiders being roughly right. | Current, and explicitly conditional: the author flags the huge ifs himself. |

| Source and locus | Claim, in its own framing | Date and status |
|---|---|---|
| S11 — individual, informal | Timelines might be medium, or broad. | Informal post, no documented evidentiary base. Useful as a live alternative, not as evidence. |

### Step 2

Steelman the long side at its strongest, which in this corpus is largely internal rather than external. Three moves do most of the work. First, the diffusion record: economy-wide transformations have taken decades even after the technology worked, and the modelling post itself uses that to justify a thirty-year slow end. Second, the incentive reading: the same source notes that continued fast progress being treated as the default is exactly the sort of extrapolative expectations dynamic that can drive bubbles, which means the update mechanism itself may be partly financial. Third, the conditional structure: the growth and automation scenarios are explicitly conditional on insiders being roughly right, and the author says so.

The crux is not whether timelines are short. It is whether the decision-relevant allocation changes, which means asking at what subjective probability reallocating grantmaker time toward short-timeline-robust work beats continuing to hedge across timelines. That is a sign-change question, and it has an answer that depends on payoff assumptions rather than on a forecast.

```
def dv(p, a, b, c):
    """Delta V(p) = V(reallocate to short-robust work) - V(hedge)."""
    return a * p - (b * p + c * (1 - p))
```

With the illustrative payoffs used below, the crossover sits near 0.67: you should only reallocate if you are roughly two-thirds confident in the short world. That number is synthetic and must be labelled as such, but the structural point is real and is exactly why the corpus itself hedges. The programme post says most of the work is worth doing under long timelines, and the CEO letter describes a deliberately diversified approach with no single theory of change. The arithmetic and the published posture agree.

### Step 3

Movers, in order of decision-relevance. First, whether the reported time-horizon measurement slows or becomes uninformative. The CEO letter contains an ambiguity worth surfacing rather than resolving: the measurement is described as increasingly difficult because systems are saturating the tasks, and saturation is consistent both with a broken measurement and with a task suite that is

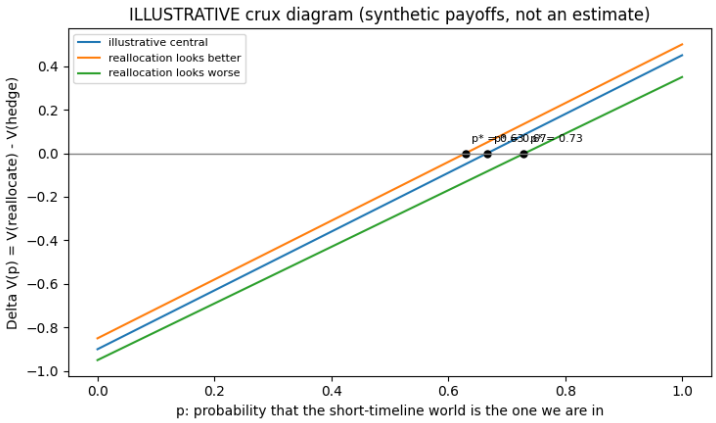


Figure 59: Graphical analysis for Step 2.

simply too easy. Naming that ambiguity is stronger than picking a side. Second, whether the capability index continues without inflection. Third, a dateable institutional test: the letter floats the possibility that AI giving does not increase dramatically in a given year even though the institution hopes it will, because it would rather prioritise longer-term capacity. If that happens, it is evidence that internal urgency translated into organisation-building rather than dollars, which is itself informative about which constraint the institution believes binds.

What would not move you, and why saying so matters: a single capability datapoint, a single staff member's update, or a restatement of the dated 2016 estimate. Each is a real signal but none is commensurate with a portfolio reallocation. The residual is that the corpus contains no distribution over staff timelines, so any statement of the form the organisation believes X is an inference from published posture, not a measurement.

The compressed answer, at interview length: the institution treats rapid progress as its operating assumption and has restructured around that, while its own leadership records that its earlier expectations kept being dashed and warns that the expectation of continuation is the kind of dynamic that can drive bubbles; the honest position is therefore to reallocate the portion of work that is robust across timelines, which is what the published posture already claims to do, and to name the measurement trend as the thing that would move me.

## Sub-Question 3: Drill B — centralised staffed grantmaking versus dispersed regrantors

Prompt: a regranting platform argues that many people's first hundred thousand dollars of grants will beat a professional grantmaker's last hundred thousand. Should a large funder regrant more, or staff up?

This is the cleanest institutional-difference question in the corpus, and it has a structure that rewards care: part of the apparent disagreement is about strategy, and part is about division of labour, and the corpus contains the arithmetic that explains the second part.

**Step 1**

Lay out the positions with their interests attached. Two of the four voices below are advocating a model they operate, and the interview answer should say so without dismissing them.

Table 141: The two claims are not symmetric. The large funder is defending a band; the small funders are claiming a niche. Both can be true simultaneously, and the interesting question is whether the boundary is set by strategy or by cost.

| Actor and rung | Position | Interest to flag |
|---|---|---|
| Coefficient Giving (P06, P01, D04) — institutional | A light staffing ratio is deliberate and measured in money moved and grants made per staff member; moving grantmaking staff onto a grantee's payroll is not necessarily a big efficiency win; regranting is impactful where there is a distinctive case, such as its work with GiveWell and a regranting challenge; first grants are preferred at the two-hundred-thousand to two-million-per-year band, and smaller projects should ideally be funded elsewhere. | Defending its own operating model; also referring smaller work to named alternative funders, which is a live division-of-labour claim rather than a dodge. |

| Actor and rung | Position | Interest to flag |
|---|---|---|
| Manifund (P11, P12) — third-party comparator | Taste and dealflow are dispersed across individuals with specialist knowledge and contacts; trust is the bottleneck to deploying a funding wave; the principal-agent problem between money and taste is structural; regranting colocates money with taste or dealflow. | **Advocates and operates the regranting model; its own hit examples are selected on the outcome.** |
| Transformative AI Fund (P10) — third-party comparator | Funds have deal flow, expertise, and operational and tooling advantages that scale and that solo donors cannot replicate; a solo donor tends to fund what is easiest to find. | Advocates the fund form it operates; note that it nevertheless argues for the same division of labour, taking the small end rather than the large. |
| BlueDot (P13, D03) — third-party comparator | Grants up to ten thousand dollars, decisions targeted at about five working days, paid upfront by default, scope deliberately widened from compute to events, community and talent acceleration. | Publishes its own grant list; its public count is not an outcome measure. |

**Step 2**

Steelman the dispersed side properly, because the naive version of it is weak. The naive version says small grants are cheap and therefore good. The strong version is an option-value argument: at the extreme small end, a grant is often the difference between a pilot existing and not existing, the diligence cost is low because the grantmaker is close to the work, and the payoff distribution is heavy-tailed, so a small number of hits pays for a portfolio. The corpus contains a concrete instance of this shape: a grant of about one hundred and forty thousand dollars in mid-2023 provided the initial funding for an organisation that later grew and merged into the entity that received what is described as the largest AI safety grant ever made, with the regranting platform computing a fifty-three-

fold growth on that single stake. Whether the multiplication is legitimate is a separate question, return to it at the mover stage.

Steelman the centralised side properly, which is easier than it looks because the corpus supplies the concession. The strongest point the large funder makes is not about efficiency ratios; it is that its best grants have been deeply infra-marginal, requiring a founder to be recruited or an organisation to be effectively incubated before there was anything fundable at all. That is the same activity the regranting argument describes, performed at a larger scale with more staff time. The fund's stated worry, that it must take ownership of problems without losing the humility of a marginal funder, is the correct frame for the whole disagreement.

Now the crux. The interesting premise is whether the boundary between the two models is a matter of taste or a matter of arithmetic. It is largely arithmetic, and the arithmetic is simple enough to do out loud: if grantmaker time is valued at an hourly rate and every grant carries a roughly fixed diligence cost plus fixed overhead for approvals, legal checks and payment, then the diligence cost per dollar falls in grant size, and a large funder will find small grants uneconomic at some threshold regardless of how good the small grant would have been.

$$\frac{\text{diligence cost}}{\text{grant}} = \frac{h \cdot r + O}{g} \quad \Rightarrow \quad g_{\min} = \frac{h \cdot r + O}{\kappa}$$

where h is investigation hours, r the loaded hourly cost of staff time, O fixed overhead, and kappa the fraction of the grant the funder is willing to spend on diligence. With illustrative parameters of twenty-five hours at two hundred dollars plus a thousand dollars of overhead, a five-percent diligence cap puts the floor at one hundred and twenty thousand dollars. That is remarkably close to the published preference floor of two hundred thousand per year, and the same arithmetic with a few hours at a very low rate puts a regrantor's floor in the low thousands, which is where the rapid-grant programmes actually operate. The bands are not a matter of taste; they are a consequence of cost structure, and the fact that each actor occupies the band its arithmetic predicts is evidence that the division of labour is working.

**Step 3**

Movers and residuals. The criterion is out-of-sample tail payoff per total dollar of cost, where cost includes diligence hours rather than only the grant. That reframing immediately disciplines the fifty-three-fold figure: it is a single case, it depends on a counterfactual (would the organisation have been seeded anyway, and would the later large grant have happened without the earlier one), and it is selected on the outcome because the platform chose it as a top hit. Attributing a

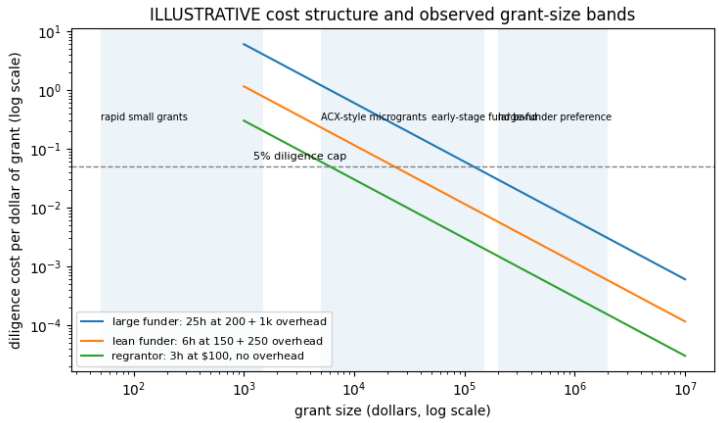


Figure 60: Graphical analysis for Step 2.

whole merger-and-large-grant chain upstream is the same error shape as treating one failed grant as a rate.

What would move you: a comparative measure of tail payoff per total dollar across the two models on out-of-sample grants; a ranked retrospective from the regranting platform's own dashboard compared against a comparable large-funder cohort; or evidence on whether the large funder's bet-big-from-the-start grants outperform its inbound-flow grants. What would not move you: an additional anecdote on either side, or a statement about total dollars moved without a counterfactual.

The residual is that no systematic comparison exists in the corpus, and both sides have an interest in the model they operate. The best available answer, stated plainly: the two models occupy different bands because their cost structures differ, the large funder should regrant where it can buy genuine option value it cannot generate in-house, and the live question is whether the boundary sits where the arithmetic says it should rather than somewhere historical accident put it.

## Sub-Question 4: Drill C — two funded papers that point in opposite directions

Prompt: your prospective employer funded two studies that appear to contradict each other on whether AI substantially increases biological risk. Which one is right?

The correct first move is to refuse the frame. This is a conditional disagreement, and the corpus says so explicitly: the organisation describes funding both, and describes supporting experts who disagree with each other as a feature rather than a bug.

### Step 1

Decompose rather than average. Two findings can be incompatible as numbers and compatible as conditional statements, and the test is whether they differ on the actor modelled, the capability modelled, the outcome measured, or the counterfactual baseline.

Table 142: Read together, the two findings bracket the chain from knowledge to harm without disagreeing about any single link.

| Finding, as reported | Actor modelled | Outcome measured | Status in this corpus |
|---|---|---|---|
| Language models could help people with little or no lab training develop new pathogens | Participants with little or no lab training | Uplift on a knowledge-elicitation task | Reported second-hand; the primary paper is not in the corpus |
| Current language models did not substantially increase the risk of a biological attack | A prospective attacker on a realistic acquisition pathway | Risk of a completed biological attack | Reported second-hand; the primary paper is not in the corpus |

```
def reconcile(a, b, dims=("actor", "capability",
                          "outcome_measured",
    "counterfactual_baseline")):
    """Classify a pair of findings as conditional or genuinely
    conflicting."""
```

### Step 2

The crux, stated as a chain: which link is binding, from hazardous knowledge through acquisition, synthesis, weaponisation and dispersal? The uplift finding bears on the first link and, if it holds, removes an assumption that expertise is the barrier. The negative finding bears on the whole chain and says the other links absorb the uplift. They conflict only if you assume the chain has one binding link.

Note that the institution's own practice is the answer to the interview question as well as to the substantive one. It funds both, it says it wants the best ideas to win over time, and it ran an essay contest whose top prize went to a paper that strongly challenged its own thinking. That is a concrete behavioural

answer for a question about handling disagreement, and it is better than a stated commitment to openness.

Mover: an end-to-end operationalisation that measures attack risk under realistic acquisition constraints while varying model access, or evidence that elicited design knowledge converts into wet-lab execution at the novice level. A related observable exists in the corpus as a grantee, not as evidence: work on automating customer screening for DNA synthesis orders is aimed at exactly the acquisition-side link that the negative finding says binds.

Residual, and a caution to state out loud: the corpus contains a summary of both papers and not the papers, so the decomposition above is a reading of someone else's summary. Related biosecurity material, such as work on mirror bacteria, is referenced without any independent assessment in this corpus, and treating a reference as an institutional view would be over-reading.

## Sub-Question 5: Drill D — does the funder accelerate capability, and what is its theory of change?

Prompt: one portfolio funds work whose stated value includes accelerating growth in developing countries after transformative AI, another funds compute and higher salaries for research organisations. Is this an acceleration funder or a safety funder, and what is its theory of change?

The single most important thing to say is that the premise is malformed: the corpus states plainly that there is no single theory of change for how to make AI go well, and describes a diversified approach with different teams organised thematically and hedging across timelines. Any answer that treats the funder as one actor is wrong on the institution's own account.

### Step 1

Answer per slice, with the locus recorded, because the sources here sit on different rungs and one of them is a staff member writing about their own grantmaking rather than an institutional position.

Table 143: Four institutional slices with different theories of change, plus one staff-level slice that is the only one explicitly about accelerating growth.

| **Slice** | **What the corpus says** | **Locus** |
|---|---|---|
| Global growth and catch-up | Grants explicitly aimed at accelerating catch-up growth after a frontier takeoff; accelerating adoption valued at roughly three hundred and eighty-one trillion in the funder's own welfare units per year of earlier adoption, approximately one hundred and fifteen million lives, conditional on the whole takeoff scenario. | Individual staff writing about their own grantmaking; the post labels its own conditions as huge ifs. |
| Technical AI safety | Request-for-proposals driven; giving roughly tripled in a year; the bottleneck is described as funding and talent rather than a shortage of useful research paradigms. | Institutional. |
| AI governance | A governance RFP closed, with capacity reallocated toward proactively identifying gaps; governance work without likely major catastrophic-risk relevance is ineligible. | Institutional. |

| Slice | What the corpus says | Locus |
|---|---|---|
| Benefits of AI | Explicit support for work that realises benefits where market failures exist, with safety framed as a form of progress; a stated wish for democracies to keep a competitive advantage. | Institutional. |
| Biosecurity | Scaled sharply across three years, with staff time described as the binding constraint. | Institutional. |

**Step 2**

Steelman the acceleration reading: funding compute for research and paying above-market salaries to keep people in safety rather than at labs is capability-adjacent; supporting work on realising benefits of AI is straightforwardly pro-capability on its own terms; and a growth-acceleration portfolio exists and says so. Steelman the safety reading: the three named risk pathways, the safeguards pillar, evaluations for dangerous capabilities, control protocols, and a dedicated team preparing for the possibility that transformative AI arrives within a few years.

The crux is whether capability or safety capacity is the binding constraint. The corpus answers with a claim rather than a hope: the bottlenecks to technical safety progress are described as funding and talent rather than paradigms, the main thing preventing the funder from scaling its impact is described as staff time, and a third-party fund's central bet is that the field lacks organisations able to absorb a funding wave. If organisational and human capacity is what binds, then funding capability-adjacent work and building safety capacity can be complements, because both increase absorptive capacity in the same ecosystem. If capability binds, they are substitutes.

Mover: absorption evidence, stated as a threshold. If significantly larger grants repeatedly clear the funder's cost-effectiveness bar, the capacity constraint is looser than claimed and the complement story weakens. If larger grants visibly degrade into funding lower-value work, the constraint binds and the complement story strengthens. What would not move you: a single large grant, or a statement of intent about scaling.

Residual: the corpus contains no portfolio-level outcome data, and the diversity that makes the question malformed also makes the answer unfalsifiable at the organisation level. The best ending is to say that the answer must be given

per fund and per decision, that the funder's own leadership names the tension between scaling existing grantees and creating new ones as something different teams will resolve differently, and that stating the malformedness of the question is itself the answer.

## Sub-Question 6: Drill E — what does the funder think? separating institution from individual

Prompt: you will be asked what Coefficient Giving thinks about a range of contested topics. The corpus mixes CEO letters, a staff member's modelling, a donor-suggestion list that explicitly disclaims institutional status, informal personal posts, other funds' launch announcements, and external maps. How do you answer without merging voices?

The answer is a habit rather than a fact: name the rung, name the date, name the interest, and say which of the four sentences you are entitled to say.

### Step 1

Build the classifier so the habit is mechanical under pressure. Each source maps to a rung, a date note, and a rule governing the sentence it licenses.

```
def use(source_id, purpose):
    """purpose in {'position', 'context', 'operation', 'forecast'}"""
```

### Step 2

Worked answers. Four claims, each with the sentence you are entitled to say.

Table 144: In three of the four rows the correction is the same shape: an individual's sentence being upgraded to an institution's sentence.

| Claim you might be asked about | Correct rung and source | The sentence you are entitled to say |
|---|---|---|
| The funder is committing over a billion dollars a year to AI safety and biosecurity | Institutional, P01 | The organisation reports that figure, and the same page notes that committed means formally recommended rather than disbursed, that including conditional approvals would raise the total, and that public databases lag. |

| Claim you might be asked about | Correct rung and source | The sentence you are entitled to say |
|---|---|---|
| Frontier AI progress will keep accelerating | Institutional post carrying a first-person update, P01 and P06 | The organisation plans on that basis, and its leadership reports having updated in that direction while warning that the expectation of continuation is itself the kind of dynamic that can drive bubbles. |
| Open weights are good, secrecy is bad, and people are too afraid of China | Individual, informal, S11 | These are one person's informal views with no documented evidentiary base in the corpus, and they sit alongside institutional statements about wanting democracies to keep a competitive advantage, so quoting them as the organisation's position would be an error in both directions. |
| GiveWell is the gold standard for individual donors | Individual staff, P07 | A named staff member's personal suggestion, on a page that states these suggestions come from individuals and teams and do not necessarily represent the institutional view. |

One case is genuinely harder and worth rehearsing: when the institution's published posture and a live internal disagreement diverge. The corpus contains an institutional statement that substantial uncertainty and expert disagreement remain, alongside dated internal estimates and a current leadership update in the opposite direction. The answer is not to pick one; it is to state the public stance, note that the disagreement is acknowledged internally, and say how you would test which view the team is currently operating under, namely by looking at what the funding decisions actually prioritise rather than at what the prose says.

Mover and residual. What would change your attribution of a claim: a byline, a

date, a fund page versus a personal post, an explicit disclaimer, or a contradicting institutional statement. What would not: the fact that many staff appear to share a view, which is a popularity signal rather than an institutional one. The residual is that the corpus gives no map of which individuals hold which views, so any statement of the form most of the team thinks is an inference and should be labelled as one.

## Sub-Question 7: Drill F — how confident are you, and what would change your mind?

Prompt: the role brief asks for reasoning that is quantitative and probabilistic, concise, calibrated, and transparent about its own limits. Expect at least one question of the form how confident are you, what would change your mind, and what do you not know about this field.

### Step 1

The institution has published its own guide to this, and the discipline it recommends is the one to demonstrate: express degrees of confidence with words or probabilities, say which considerations are most important, and say what kind of support you have for a claim. It also notes the trap, namely that words like plausible are interpreted differently by different readers.

The rule to follow under interview pressure: convert only when the source converted. If the source said seems likely, you may report that it means above fifty percent with no detailed investigation. If the source said plausible, you may not report a number at all. The conversion of a qualitative hedge into a two-digit probability is the most common form of false precision in this domain, and it is easy to catch when someone is listening for it.

```
def encode(term, as_number=False):
    """Refuse to launder a qualitative hedge into a number."""
```

### Step 2

The second discipline is the criterion. What would change your mind only counts as an answer if the criterion can come out against you. The writing-structure source states this directly for the methodological move: asking the other side to produce an impossibly complete alternative makes it too easy for your own position to survive. Watch for this failure in your own answers and name it if you catch yourself doing it.

A corpus-anchored answer to what would change your mind about an institutional practice: the institution demonstrates the behaviour itself. One grantmaker describes the habit of asking early and often which piece of evidence would most likely change his mind, and seeking it out, including booking a

short call with the expert who could move him. The organisation ran an essay contest for views that would influence its thinking and gave the top prize to a paper that strongly challenged it. Both are better answers than a stated commitment to updating.

```
def check_criterion(observable, direction, would_falsify):
    """Reject criteria that cannot come out against the position."""
```

**Step 3**

The third question is what do you not know, and the corpus is unusually good at supplying the shape of an honest answer, because several sources publish their own omissions.

- A governance map lays out seventeen activities and three pathways and then names five macro-activities it deliberately leaves out, and adds a jurisdictional disclaimer that its account is derived from work in the EU, the US, the OECD, France and the G7, and will not generalise, giving advocacy in China as an example.
- A policy retrospective reports that almost all its proposals were set aside for political infeasibility, and explains a failed attempt to build an IPCC-style body for biotechnological risk in terms that generalise: no consensus among experts, misaligned incentives because raising the alarm would restrict the researchers' own field, and no champion willing to spend political capital. It also notes the mundane constraint that a parliamentary committee cannot do original research, and that policymaker agreement that a problem exists does not produce anyone who will develop responses.
- A grant round reports six hundred and twelve unfunded applications out of six hundred and fifty-four, and decomposes the reasons in prose without any counts, which is a clean example of a decomposition you can describe but not yet weight.
- A programme publishes its own open questions, including how much to value different kinds of career shift, where diminishing returns bite, whether it is being drawn toward what is easier to measure, and how much it should shape an ecosystem versus respond to it.

Close with the base-rate caution, because it is the discipline that separates a good answer from a persuasive one. A single failure or success story demonstrates that a mechanism exists; it does not estimate how often the mechanism fires, cannot rank mechanisms by frequency, and does not justify a process change on the grounds that it happened once. The same caution applies in the other direction: treating every unfunded application as a false negative is the mirror-image error.

## Sub-Question 8: Capstone — the ninety-minute rehearsal circuit and its limits

Run the seven drills in a single sitting, on a clock, out loud. The circuit below budgets about seventy minutes of work plus setup and debrief, which fits the ninety-minute slot. Round one is a solo answer followed by a card and a self-score. Round two repeats each question in the opposite persona: hold the steelman position as if it were yours, which is the fastest way to find out whether your steelman was actually strong.

Score against eight criteria, two of which carry extra weight because they are the ones most often faked: whether the steelman is the opponent's strongest version rather than a strawman, and whether the crux comes with an observable.

### Step 1

The circuit, with times. Each drill gets ninety seconds to answer, sixty seconds to self-score, and four minutes to write the card. The adversarial replay gets two minutes per drill and the debrief gets ten.

```
def circuit(drills, phases, replay_s=120, debrief_s=600):
    """Total minutes for the full rehearsal circuit."""
```

### Step 2

The rubric. Score each drill from zero to two on each criterion. One rule is a hard fail rather than a deduction: attributing an uncited claim to the institution.

Table 145: Hard fail: an uncited claim attributed to the institution. That row scores zero for the drill regardless of everything else.

| Criterion | What a 2 looks like | Weight |
|---|---|---|
| Locus tagged | Every cited source carries a rung and a date note, and no individual view is upgraded to an institutional one. | 1.0 |
| Position scoped | Actor, capability, outcome, baseline and the relevant decision are all pinned down before the argument starts. | 1.0 |

| Criterion | What a 2 looks like | Weight |
|---|---|---|
| Steelman is the strongest version | The opposing case is stated in terms its own holder would recognise and endorse, including the parts that are uncomfortable. | 1.5 |
| Crux has a criterion and an observable | The premise is stated so that a sign change is possible, and the measurement that would resolve it is named. | 1.5 |
| Mover has a threshold | A direction and a magnitude are given, and the answer says what would not move you. | 1.0 |
| Residual names the gap | The answer ends with the specific thing the corpus cannot settle and what you would ask for. | 1.0 |
| No numeric laundering | Qualitative hedges stay qualitative, and every number carries its source and date. | 1.0 |
| Hypotheticals labelled | Every synthetic parameter, invented scenario and illustrative figure is flagged as such at the moment it is used. | 1.0 |

**Step 3**

What this set can and cannot do, stated plainly so that it is never overclaimed in the room. It can train the form, the attribution discipline, the steelman, the crux test and the calibration habits, against sourced positions rather than invented ones. It cannot supply interviewer signal about what the panel weights, it cannot supply the primary papers behind the institutional biorisk summary, and it cannot turn a small number of illustrative anecdotes into a rate. It also cannot substitute for the one thing the corpus repeatedly says is the actual work: going and finding the opportunity that would not have arrived on its own.

Two closing habits worth carrying into the room. First, when a disagreement

appears, ask which rung it is on before asking who is right, because many apparent conflicts in this corpus dissolve at the rung boundary rather than at the level of fact. Second, when you are asked what you think, answer with the structure rather than the verdict: the position, the strongest opposing case in its own terms, the premise that decides, the observable that would move you, and the gap you cannot close from here. That structure is the deliverable, and it is the thing being hired.

# Write the final one-page positioning note and source index.

## Question

Produce the closing artefact of the preparation pack: a one-page interview positioning note for an evaluator or grantmaker role, plus a concise indexed source map. The note must contain a roughly 200-word answer to the question **why this role, and how would you approach it**.

*Hard constraints. The note is one page. Nothing may be invented: no application, no credentials, no work history, no experience. Every plan is marked as a plan, every inference as an inference, and a place where a track record would normally sit either holds a slot the learner fills from their own real history or states plainly that the episode has not happened yet. The index points to source labels and file names rather than reproducing corpus text.*

Separation requirement. The index keeps four categories apart and marked: role requirements, source-supported claims, proposed exercises, and personal inferences. A reader should be able to tell, without re-reading the corpus, which line would survive an attribution challenge and which line is the learner's own construction.

## Background

This is the last step of the final phase, and it depends on the phase's earlier outputs rather than on new source material: the crux table tagged with the four-level attribution ladder, the funder map parameterised by mandate, capital, evaluation cost and risk posture, the uncertainty register with its ten fields, and the failure-mode catalogue with its censoring and selection caveats. The conversation history for this step is empty, so no new ordering constraint has been imposed and the note is assembled from what the phase already established.

Three source families set the standards the artefact has to meet. D01 gives the transparency standard: open with a linked summary of key takeaways, indicate which considerations are decisive, and state confidence together with the kind of support behind each major claim. L02 supplies the reader-side argument structure: meet the reader where they are, take them to the claim, show what could move you, show the evidence, show whether it moved you, and show what should now change for them. L01 supplies the budget and project-management vocabulary (work breakdown, cost breakdown, cost drivers, tiers, ownership) that makes a proposal legible to an external reviewer.

The role brief fixes the assessment surface the note must exhibit: comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, judging people and organisations, and communicating concisely, calibratedly

and transparently. A one-page note is the smallest artefact that can display all of these at once, which is why it closes the pack.

## Sub-Question 1: Draft the positioning note and the roughly 200-word answer

Write the note itself, including the roughly 200-word answer to why this role and how you would approach it. The answer is the load-bearing sentence block of the whole pack: it is where a candidate normally reaches for a track record, and it is therefore where the no-fabrication constraint is most likely to fail quietly.

Success condition: a reader who knows the corpus can verify every substantive claim in the note against a source label, can identify the one bracketed slot left for real personal history, and cannot mistake a plan for a demonstrated capability.

### Step 1

Fix the skeleton before writing prose. L02's account of the introduction and conclusion is the governing analogy: they are written once the argument exists, because until then the writer does not know which contribution deserves attention. Here the sequence of sections is the argument, and fixing it first prevents the note from becoming a biography with an appendix.

- Frame (role requirement). What the role is, in the brief's own terms: grantmaking across AI governance, catastrophic-risk capacity building, short-timeline projects and biosecurity; investigating opportunities, developing neglected projects, recruiting founders, launching initiatives, researching programme strategy and historical cost-effectiveness, and maintaining grantee and stakeholder relationships. Anchor: the role brief.
- Why this role (source-supported reasons, not biography). Two or three reasons tied to named material: the three pillars of visibility, safeguards and capacity (P08); the AI-governance subject areas the fund actually solicits (D04); the seeding-versus-scaling tension the CEO names as the live institutional problem (P06).
- The answer (roughly 200 words). Why this role and how you would approach it, drafted in the next step with its slots visible.
- Method. The six criteria D04 states (theory of change, track record, strategic judgment, project risks, cost-effectiveness, scale), the crux-first discipline, and the mind-changing-evidence question from P04.
- Boundary and uncertainty. What is planned rather than demonstrated, which timeline estimate is dated rather than current, and which conclusions the note deliberately does not resolve.

- Close. A 90-day plan stated as three observable outputs, not as an ambition.

**Step 2**

Write the answer with explicit slots rather than implied biography. Any position where an achievement would normally sit is either a bracketed slot the learner fills from their own history, or an explicit statement that the episode has not happened. Marking convention for the rest of the note: proposed items are prefixed **Proposed:** and the learner's own judgments are prefixed **Inference:**.

I am drawn to this role because it is the job of deciding where money is most usefully placed against catastrophic risk, and of owning that decision rather than advising on it. On substance I would work from what Coefficient's own materials treat as settled: the three pillars of visibility, safeguards and capacity (P08), the six AI-governance subject areas the fund solicits (D04), and the biosecurity bottlenecks named in the four-pillar plan — protective equipment, biohardening, pathogen-agnostic detection, and rapid medical countermeasures (P06).

On method I would start from the six criteria D04 states, then use the habit Trevor Levin describes: name the single piece of evidence most likely to change my mind, and find the cheapest reliable way to obtain it (P04). I would write the investigation the way D01 asks — key takeaways first, the decisive considerations flagged, my confidence and the kind of support stated, sources linked — because that is the form in which someone else can check my reasoning without repeating my work.

***[Slot: your own concrete episode — what you investigated, what you recommended, what you got wrong. Fill from real history. If no such episode exists, the honest sentence is 'I have not yet carried a grant investigation', followed by what you would do to earn the capability.]*** I would rather say that plainly than dress adjacent experience up as grantmaking. In a first 90 days I would learn one portfolio's bar, take three well-scoped opportunities from first contact to written recommendation, and produce one memo a colleague could act on without me in the room.

Proposed: this paragraph is a template, not a transcript. The bracketed slot is the only place a track record may appear, and it is legitimate for it to read 'not yet'. Run the step code to test the draft mechanically.

What the check does. It splits the draft into sentences, skips any sentence containing a bracketed slot or an explicit hedge, and flags the remainder if it contains an achievement verb. The flag list is a review queue, not a verdict: the point is that every flagged sentence forces an explicit decision about whether the claim is real, planned, or absent. It also reports the word count so the roughly 200-word target is verified rather than estimated.

**Step 3**

Write the method paragraph so it is checkable rather than admiring. The role brief asks for comparison of opposing arguments, quantitative and probabilistic reasoning about expected value, identification of decisive considerations, and proportionate due diligence. A method paragraph that lists virtues demonstrates none of them. The paragraph below performs the moves instead.

For any opportunity I would first locate the crux: the premise that, if it flipped, would flip my recommendation. I would then name the observable that would flip it, accept the cheapest evidence that bears on that observable, and stop diligencing at the point where the next hour of work could not move the decision. Where two funded sources disagree, I would not average them. The stated example is the disagreement over whether current language models materially increase biological attack risk, where the corpus itself notes that a finding of uplift and a finding of no substantial aggregate risk can be compatible once the actor modelled, the capability modelled, and the counterfactual baseline are separated (P08).

*Size and posture follow from the same logic, not from taste. On a first-time grant D04 states a preference for roughly $200k to $2m per year over one to two years, with renewal considered once there is evidence of success; on an early-stage bet, the third-party comparators in the corpus describe bands an order of magnitude smaller, and the point of those bands is tolerance of failure probability rather than enthusiasm for small numbers. Inference: my default would be to match the band to the evidence available, and to say out loud which band I am in and why.*

Proposed: commit to writing the recommendation in D01's order — takeaways first, decisive consideration flagged, confidence and support kind stated — and to naming at least one consideration that would have changed the answer and did not.

**Step 4**

Write the boundary and the close. A positioning note that concedes nothing reads as either unaware or dishonest; a note that concedes the right things reads as calibrated. The concessions must be chosen, not scattered, and each must be a concession whose resolution the learner can name.

- Capability boundary. No grant carried through this funder's process, stated as such, with the compensating plan attached. This is the only concession about the person.
- *Dating boundary. The only quantified institutional timeline claim in the corpus is a 2016 estimate of at least a 10 percent chance of transformative AI by 2036; the current institutional posture is qualitative (P01) or directional (P06), and one individual staff voice argues timelines might*

*be medium or broad (S11). Never present the 2016 figure as a present forecast.*

- *Conditionality boundary. The catch-up-growth material is a staff member's model that conditions on insiders being roughly right, assumes automation of most physical tasks, and carries the author's own list of ways it could be wrong (P03). Cite it as conditional or not at all.*
- Deliberate non-resolution. The timeline crux is held open rather than settled, because the corpus does not supply the evidence that would move it; the note should say what would.

Close on three observable outputs rather than a sentiment: one portfolio's bar learned well enough to state it back, three proportionate investigations completed end to end, and one written recommendation another person could act on without the author present. Each output is checkable by a third party, which is the property that an interview answer about the first 90 days usually lacks.

## Sub-Question 2: Build the concise source index and the four-way separation map

Produce the index that sits beside the note: a compact map from source label to what the source is, its attribution level, and the one use made of it in the note. Then produce the separation map that keeps role requirements, source-supported claims, proposed exercises and personal inferences visibly distinct.

*Constraint: the index points at original files and labels, does not reproduce corpus text, and stays short enough to sit on the same page as the note.*

### Step 1

Fix the attribution legend before writing any row. The phase established four levels, and the index is unreadable without them because the same sentence can be institutional in one source and a named person's view in another.

- **Institutional. A published fund or organisation position. Marks as [INST]. Example family: P01, P06, P08, P05, D01, D04. Usable as the organisation's stance, subject to a date check.**
- **Individual staff. A named person's model, suggestion or opinion. Marks as [IND]. Example family: P03, P07, S09, S10, S11. Never attributable to the institution.**
- **Third-party comparator. An external fund or programme with its own incentives. Marks as [3P], with a self-interest tag where the source argues for a model it operates.**
- **External primary and local authored material. Material cited by the funder but authored elsewhere, plus the locally authored method documents. Marks as [EXT] and [LOCAL] respectively.**

*Reading rule carried over from the phase and applied here: prefer the source closest to the decision and least distorted by incentives, but use a self-interested source freely for operational facts it asserts about itself, such as its own grant sizes, timelines and process.*

**Step 2**

Write the source map. One row per source family, one line of use, no reproduction of corpus text. Point to the file names supplied with the brief: the role brief, the 41-record literature file, and the local companion documents.

Table 146: Source map for the one-page note. Labels match the brief's own prefixes; levels are the phase's attribution ladder.

| Source | What it is | Level | Use in the note |
|---|---|---|---|
| `evaluator-job-ad-Anthropic-2026-09-16.txt` | Anthropic 2026 role brief: scope, required dispositions, learning intent, no-fabrication instruction | [INST] | The frame section; the assessment surface the note must exhibit |
| `P01` | Oehlsen on scaling AI and biosecurity giving; Project Tailwind | [INST] | Direction of travel, urgency claim, grant size and speed changes |
| `P04` | Day in the life of three GCR grantmakers | [INST] staff-voiced | The mind-changing-evidence habit; grant investigation as the unit of work |
| `P06` | CEO letter: scale, seeding versus scaling, marginal versus inframarginal | [INST] | The live institutional tensions used to make the why-this-role reasons specific |

| Source | What it is | Level | Use in the note |
|---|---|---|---|
| P08 | Approach to AI safety and security | [INST], pre-rebrand and dated | Three pillars; the dated 2016 timeline estimate; the biorisk disagreement example |
| D01 | Reasoning transparency | [INST] method document | The writing standard the note and any memo must meet |
| D04 | AI governance RFP, now closed | [INST], closed-process detail may date | Six assessment criteria; the $200k-$2m per year first-grant band; EOI-first process |
| P10, P11, P12 | EA Funds Transformative AI Fund; Manifund regranting; Manifund on grant-program design | [3P], self-interested | Contrasting fund designs and size bands; cite for the design claim, tag the self-interest |
| P02, P05, P07 | Careers RFP reflection; 2026 careers strategy; per-staff donor suggestions | [INST] plus [IND] in P07 | RFP design lessons; the explicit statement that per-staff suggestions are not the institution's view |
| P03 | Catch-up growth after AI | [IND] named staff author | Cited only as a conditional model carrying its author's own error list |

| Source | What it is | Level | Use in the note |
|---|---|---|---|
| S09, S10, S11 | Oehlsen on Project Tailwind; Chen on funding pitfalls and on disagreements | [IND] | Individual disagreement held as disagreement; not merged into institutional voice |
| L01, L02 | Local method documents: budget and project structure; argument and reader structure | [LOCAL] | How to cost a proposal; how to order the note so a reader can follow it |
| N01–N06, N14, S01–S08 | Comparator applications and grants; governance map; policy-feasibility warning | [EXT] and [3P] | Territory map for the frame; feasibility caveat; worked comparator material |

**Step 3**

Write the separation map. This is the table that answers the question the deliverable turns on: which sentence would survive an attribution challenge, and which is the learner's construction. It is also the table that keeps a rehearsal answer honest under questioning, because the candidate can say which category a line belongs to without hesitating.

Table 147: The four categories the deliverable requires the index to separate.

| Category | Examples in this note | Marking |
|---|---|---|
| **Role requirements** | The six D04 criteria; the EOI-first process; ownership of ambiguous work; concise and transparent communication | Quoted or paraphrased with the source label, no interpretation added |

| Category | Examples in this note | Marking |
|---|---|---|
| **Source-supported claims** | Three-pillar structure; grant-size bands; giving figures; the biorisk reconciliation; the four-pillar biosecurity plan | Source label plus level; dated items flagged as dated |
| **Proposed exercises** | The 200-word answer; the crux-first memo; the 90-day plan; the mechanical draft check | Prefixed Proposed; never presented as something already done |
| **Personal inferences** | Where the learner would fit; the ranking of reasons; the default band-matching heuristic | Prefixed Inference; carries a confidence word and a named mover |

**Step 4**

Compress. The index is a pointer, not a précis: labels, levels, and one clause of use per row. Where two sources make the same point, keep the one closer to the decision and drop the other. Where a source is cited only for an operational fact about itself, say so in the use column rather than awarding it evidentiary weight it does not have.

- Cap the map at roughly fifteen rows and group the comparator material into single rows.
- Never reproduce corpus prose; a label plus level plus use clause is sufficient.
- *Tag self-interest on any source that argues for a model it operates, so the reader can discount it without being told to.*
- Keep the level tag identical to the phase's ladder vocabulary, so the note and the index can be checked against each other.

## Sub-Question 3: Run the fabrication and status check, then freeze the one-pager

*Test the assembled artefact against the constraint that governs the whole pack, then freeze it. The test is not stylistic: every failure mode here produces a document that reads better than it deserves, which is precisely the failure an evaluator is trained to catch.*

Also fix the physical constraint: the note, its answer, and the index must fit on one page, which forces a word budget and therefore a decision about what the note does not attempt.

**Step 1**

Run the eight checks. Each check exists because a specific, plausible drafting failure produces a document that looks stronger than the evidence behind it.

1. Credential check. Any sentence implying a capability not yet demonstrated passes the step code as a bracketed slot or an explicit not-yet, or it is rewritten. Catches: the implied biography.
2. Dating check. Every quantified institutional claim carries a date, and the only dated timeline figure in the corpus is flagged as historical rather than current. Catches: numeric laundering of an old estimate into a present forecast.
3. Voice check. No individual staff view appears as an institutional position, and the per-staff nature of the donor-suggestion material is preserved. Catches: voice merging.
4. Crux check. Every disagreement named in the note carries a clause and a counter-holder, not just a topic. Catches: disagreement as gesture.
5. Comparator check. Third-party sources are tagged, and self-interest is flagged wherever the source advocates a model it operates. Catches: treating an interested comparator as neutral.
6. Prefix check. Every proposed item carries Proposed and every personal judgment carries Inference, with a confidence word and a named mover. Catches: status flattening.
7. Conditionality check. The catch-up-growth material is either cited with its condition and its author's error list or is absent. Catches: quoting a conditional model as a finding.
8. Evidence-kind check. Every major claim in the note says what kind of support stands behind it in D01's vocabulary, from a careful examination of primary material down to a general intuition. Catches: a claim whose support kind the candidate could not name if asked.

**Step 2**

Fix the budget. A one-page note that must contain a roughly 200-word answer plus an index has room for about 550 to 650 words of body text. The budget below is the allocation, and it is the reason the note can be honest: something has to be left out, and the note says what.

Table 148: Word budget for the one-page note: roughly 490-580 words of body text plus the index.

| Section | Target | Must contain | Anchor |
|---|---|---|---|
| Frame | 70-90 words | The role's own scope plus the pillar structure | role brief; P08 |
| Why this role | 80-100 words | Two or three reasons tied to corpus material, not biography | P06; P08; D04 |
| The answer | roughly 200 words | The why-and-how text with its bracketed slot visible | P04; D01 |
| Method | 60-80 words | Six criteria, crux-first, mind-changing evidence, stop rule | D04; P04 |
| Boundary and uncertainty | 50-70 words | No fabricated track record; dated anchor; held-open crux | P08; S11; P03 |
| Close | 30-40 words | Three observable 90-day outputs | L01; L02 |

**Step 3**

Rehearse the note aloud, because the note is not the deliverable; the ability to produce it under questioning is. Use L02's fourth exercise: give a two-minute argument, then ask the listener to restate the claim, what could change it, and what follows. Their reconstruction is the diagnostic. If they can state the claim but not the mover, the method paragraph is doing less work than it appears to.

*Freeze the document with a deliberate omission list, so the omissions are visible rather than accidental: no biography, no invented episodes, no third-party claim presented as the funder's own position, and no resolution of the timeline crux. State in the note that the crux is held open and what would close it; an unresolved crux handled explicitly reads as calibrated, while the same crux handled silently reads as evasion.*

Final status of the artefact. The one-page note and the index are complete when a reader can verify every source-supported line against a labelled source, identify the single slot reserved for real personal history, see which items are

proposals rather than achievements, and name the one question the candidate would ask of the funder rather than answer for it. That question, formed but left unstated in the note itself, is the natural close of the pack.

# Run a mock interview and feedback loop.

## Question

Run a self-administered mock interview of **45 to 60 minutes** for the evaluator or grantmaker role, capture a transcript, score it against the role criteria, produce a specific revision list, and then run a second pass on the revised answers.

The deliverable is the loop, not the transcript: a scored first pass, a ranked revision list justified by expected marginal gain per minute, a second pass on fresh prompts of the same type, and a written list of limitations the loop cannot repair. Nothing personal may be invented: where an item invites a lived example, answer with a clearly labelled hypothetical plus an honest statement of what is missing.

## Background

The role brief states the evaluation criteria that must become elicitable behaviours: comparing opposing arguments, reasoning quantitatively and probabilistically about expected value, identifying decisive considerations, choosing proportionate due diligence, judging people and organisations, and communicating concisely, calibrated and transparently about reasoning. The supplied corpus adds one more dimension the brief implies but does not name: **provenance discipline**, because the corpus contains institutional positions, individual-staff views, self-interested comparators and external primary material that must not be merged.

Source anchors for the instrument: D04 gives the assessment criteria used in a real Coefficient Giving RFP, including theory of change, track record, strategic judgment, project risks, cost-effectiveness and scale, and states that reasoning transparency and scope sensitivity are valued. P04 describes the working habits of grantmakers, including asking early what piece of evidence would be most likely to change one's mind and seeking the most decision-relevant information as time-efficiently as possible. D01 gives the reasoning-transparency standard. L02 gives the structural standard for a written or spoken argument.

Counterargument and feasibility pressure should be sourced, not invented: P03 supplies a conditional disagreement about catch-up growth and labels its own assumptions as large ifs, and N14 supplies the case that sound technical proposals fail without political feasibility, coalitions and implementation work. These are useful precisely because they resist a tidy answer.

Stated uncertainty: the corpus does not establish the real interview format. It may be live, written, or include a work sample or case exercise, so both a spoken and a written variant are prepared. A second unknown is the fraction of resolved checkable claims in a transcript, which limits how strong any calibration score can be.

## Sub-Question 1: Subquestion 1: Build the instrument and verify coverage

Design the mock as a measurement device. Each role criterion must be elicited by a task that forces the behaviour, and each task must have an observable signature and a failure signature, so that scoring can cite the transcript rather than an impression. Then verify that the set covers every criterion and that no criterion is measured by only one item.

### Step 1

The first non-trivial move is refusing items that can be answered by reciting a source. A question such as what does the corpus say about timelines measures recall. A question such as which premise would have to flip before you would reallocate grantmaker time measures judgment. Each row below pairs a criterion with the item type that forces it, the observable that confirms it, and the observable that disconfirms it.

Table 149: Instrument map from role criterion to eliciting item, observable signature and failure signature.

| Criterion | Eliciting item type | Observable signature | Failure signature |
|---|---|---|---|
| Comparing opposing arguments | Give two live positions on a contested claim and ask for the strongest version of each | States both positions, names the variable that discriminates them, states which side wins under which value | Picks a side first and caricatures the other, or splits the difference into a mean |
| Quantitative and probabilistic expected value | Ask for a back-of-envelope expected value with a probability and a payoff | States p, states R, shows the product, shows sensitivity to each input | Reports a point estimate with no probability, or uses a benefit number with no baseline |
| Identifying decisive considerations | Ask which premise, if flipped, would change the decision | Names one premise and the direction of the flip that reverses the decision | Names a topic or theme rather than a variable |
| Proportionate due diligence | Ask what the next 30 minutes of investigation should buy | Names the cheapest decisive evidence, its cost, and a stop rule | Proposes maximal diligence, or decides on no evidence |

| Criterion | Eliciting item type | Observable signature | Failure signature |
|---|---|---|---|
| Judging people and organisations | Ask for a read on a team from a thin evidence package | Names a behavioural indicator, a base rate, and the evidence that would overturn the read | Relies on credentials or tone, or refuses to judge at all |
| Communication: concise, calibrated, transparent | Ask for a 90-second recommendation with an explicit confidence | One claim, one reason, one uncertainty, one number | Lists everything, or hedges without ever committing to a number |
| Provenance and attribution | Ask which part of an answer is source-supported, which is inference, which is exercise | Separates the three and names the attribution level of each source | Speaks as the institution while quoting individual staff, or treats a self-interested comparator as neutral |

The second non-trivial move is the adversarial injection. The interviewer must interrupt at least once with a strong counterargument or a new piece of evidence and then score the ***update***, not the original answer. Without an injection, openness to counterevidence and calibration are unmeasured, because a static transcript cannot distinguish a candidate who updates from one who happened to be right. P04 supplies the discipline: ask early what evidence would change your mind, then seek it.

- Coverage floor: every criterion must be elicited by at least two items, because a single-item measurement cannot distinguish a stable capability from a lucky answer.
- Mixed items are allowed and realistic, but the score must be decomposed into the sub-criteria they hit; credit given to a mixed item may not be counted twice in the aggregate.
- At least one item must make the correct move an explicit refusal to decide plus the cheapest evidence that would resolve the question. This is the item that separates decisive from rash, and it is also the item that measures proportionate due diligence.
- At least one item must supply an obviously settled-looking case in which the right answer is that the evidence is thinner than it appears, to test whether the candidate rewards confident presentation.

- Every item carries a fixed time budget. Time overrun is itself a scored observation, because P01 treats decision speed as part of the job rather than an afterthought.

### Step 2

The coverage check is small but load-bearing. It converts a claim of coverage into a structure that can fail: uncovered criteria, criteria measured by a single item, and items that hit no criterion at all.

```
items = input_state["items"]          # {item_id: set(criterion_ids)}
criteria = input_state["criteria"]    # list of criterion_ids

covered = {c: [] for c in criteria}
for iid, cs in items.items():
    for c in cs:
        covered.setdefault(c, []).append(iid)

uncovered = [c for c in criteria if not covered.get(c)]
single_item = [c for c in criteria if len(covered.get(c, [])) == 1]
orphan_items = [i for i, cs in items.items() if len(cs) == 0]

# a criterion hit by one item is fragile; report the item so it can
    be duplicated
fragile_pairs = {c: covered[c] for c in single_item}

output_state["uncovered_criteria"] = uncovered
output_state["fragile_pairs"] = fragile_pairs
output_state["orphan_items"] = orphan_items
output_state["coverage_report"] = covered
```

*Caveat on the check itself: coverage is a property of the item set, not of the answers. A perfectly covered instrument still yields weak evidence if the answers are scored from memory rather than from cited transcript lines. The check removes one failure mode and creates no guarantee about the others.*

## Sub-Question 2: Subquestion 2: Run the first pass and produce the transcript

Run the six-item set under the clock and write the answers verbatim. The transcript is the data; summaries written afterwards cannot be scored. Every answer that sounds autobiographical is a labelled hypothetical, stated in the conditional, with the missing ingredient named explicitly.

### Step 1

The item set below is one instantiation of the instrument. It is designed so that the hard part of each item is the reasoning move, not the recall of corpus facts.

1. Grant evaluation. Two opportunities on the table: Opportunity A, a two-person team with a working prototype asking $60k for a six-month AI governance research pilot with no institutional home and one BlueDot alum; Opportunity B, an established 501(c)(3) with a three-year track record asking $1.5m per year for two years to scale a safety evaluation program. Recommend one, with a probability and the decisive consideration.
2. Opposing arguments. Two live positions on whether shortened timelines should reallocate grantmaker time now toward short-timeline-robust work. Give the strongest version of each, name the discriminating variable, and state the condition under which each wins.
3. Expected value. Give a back-of-envelope expected value for a $60k pilot with a stated hit probability and a stated payoff, then show the sensitivity to each input separately.
4. Proportionate due diligence. State what the next 30 minutes of investigation should buy for Opportunity A, what it costs the applicant, and the point at which further diligence would be disproportionate.
5. Judging people and organisations. From a thin package, give a read on the two-person team. Name the behavioural indicator you are reading, the base rate you are comparing it to, and the evidence that would overturn the read.
6. Communication. Give a 90-second recommendation on whether to fund a $10k rapid grant to a national community organiser, with an explicit confidence and one sentence on what would change it.

Labelling rule, stated before the clock starts: any answer that would sound like a personal history is delivered as an explicit hypothetical, for example, if this case were in front of me I would first ask what the pilot's primary deliverable is. The trap to avoid is not lying but ***borrowing***, which is narrating in the first person an outcome produced by a grantee, a colleague, or a corpus author. That is scored as an integrity gate failure, not as a stylistic slip.

**Step 2**

Two moves in this step are non-trivial. The first is that the diligence question is an expected-value-of-information question, not a thoroughness question, so proportionality has a stopping condition that can be computed.

$$\max_a \mathbb{E}\big[V(a)\big] \qquad \text{investigate while } \mathrm{EVI}(e) \geq c(e)$$

$$\mathrm{EVI}(e) = \mathbb{E}\big[\max_a V(a \mid e)\big] -$$

The second is the compression rule. The same content is prepared at three lengths, and the cut removes elaboration rather than the decisive consideration

or the uncertainty. Cutting the uncertainty statement is the most tempting and most damaging edit, because it changes what the answer asserts.

Table 150: Skeleton of the first-pass transcript. The decisive consideration and the uncertainty are the two columns that must survive compression.

| **Item** | **One-sentence claim** | **Decisive consideration named** | **Uncertainty stated** | **Sources used and their level** |
|---|---|---|---|---|
| A. Grant evaluation | Fund the pilot at a small size or refer it to an early-stage funder rather than fund the scale-up | Deployable absorptive capacity per grantmaker hour, not the size of the request | Stated probability that the pilot produces a usable governance output | D04 institutional on scale bands; P10 third-party comparator on small-grant niche |
| B. Opposing arguments | Reallocate only after a stated trigger, because the two positions are conditional rather than contradictory | Whether capability evidence is accelerating or stalling, not how alarming the scenario feels | Names the observable and says the corpus does not settle it | P01 and P06 institutional; P08 institutional but dated; S11 individual staff |
| C. Expected value | The pilot is justified if p times R exceeds the cost, and the decision is more sensitive to p than to R | The p estimate, because a plausible range crosses the threshold | States the range for p and the resulting range in EV | D02 institutional on BOTEC practice |

| Item | One-sentence claim | Decisive consideration named | Uncertainty stated | Sources used and their level |
|---|---|---|---|---|
| D. Due diligence | One fifteen-minute call with the team's technical reference buys more than a week of reading | Whether the claimed prototype has an internal user | States what would make the call unnecessary | P04 institutional on decision-relevant evidence and avoiding unnecessary work for applicants |
| E. Judging people | Weight shipped artifacts over stated intentions, and one shipped artifact over credentials | Whether the artifact was used by anyone other than its author | Names the base rate problem: thin packages over-select confident presenters | P06 institutional on early bets that did not look propitious |
| F. Communication | Fund at the smallest size that tests the organiser's rate of shipped events | Event throughput per dollar, which is observable within one quarter | Explicit confidence and one sentence on what would change it | D03 external primary on the rapid-grant ceiling and decision latency |

## Sub-Question 3: Subquestion 3: Score the first pass with gates before dimensions

Score the transcript. Separate binary integrity and decisiveness gates from continuous quality dimensions, then compute a calibration measure over the claims the transcript actually commits to. The output is a revision list, not a grade.

### Step 1

For this rehearsal, use editorial binary integrity checks alongside continuous quality dimensions so that strong presentation does not conceal invented credentials or attribution errors. This scoring design is a practice choice, not a validated instrument or a claimed Coefficient Giving hiring rule. A justified decision to seek decisive evidence can itself be a good answer.

Table 151: Binary gates. A failed gate caps the overall result and cannot be averaged away.

| **Gate** | **Pass condition** | **Evidence line required** |
|---|---|---|
| G1 No invented credentials | Every autobiographical-sounding passage is labelled as hypothetical | Quote the passage and its label, or note the absence of a label |
| G2 No borrowing | No first-person claim of an outcome produced by another person or organisation | Quote the first-person outcome claim and name the actual producer |
| G3 Uncertainty stated | Material uncertainty is stated with appropriate confidence words or a probability/range; use numbers when the exercise calls for them and the number has an interpretable basis. | Quote the uncertainty statement and assess its basis; do not penalize a qualitative statement merely for lacking a number. |
| G4 Decisive consideration named | When asked for the decisive consideration, a variable is named rather than a topic | Quote the named variable and the flip direction |
| G5 Attribution discipline | Institutional, individual-staff, third-party comparator and external primary sources are distinguished where relevant | Quote two attributed claims and check the level assigned to each |

Table 152: Continuous dimensions. Anchors are described at 1, 3 and 4; interpolation is permitted.

| **Dimension** | **Anchor at 1** | **Anchor at 3** | **Anchor at 4** |
|---|---|---|---|
| D1 Clarity and structure | Reader cannot reconstruct the claim | Claim, reason and uncertainty are recoverable | Each move prepares the next, in the L02 sense |

| Dimension | Anchor at 1 | Anchor at 3 | Anchor at 4 |
|---|---|---|---|
| D2 Calibration | Stated confidences are systematically too high | Resolved claims land inside the stated range | Ranges are narrow where evidence is strong and wide where it is thin |
| D3 Reasoning transparency | Conclusions asserted without support kind | Most important considerations flagged, support kind indicated | Key takeaways opened with, confidence and support given per major claim |
| D4 Decisiveness and speed | No recommendation offered | Recommendation given inside the time budget | Recommendation given inside budget with a named stop rule for reversibility |
| D5 Openness to counterevidence | Repeats the prior answer after the injection | States what would change the mind and updates when it arrives | Names the update, its size, and the residual uncertainty that remains |
| D6 Proportionality | Diligence is either maximal or absent | Diligence is scaled to stakes and reversibility with a stop rule | Compares the cost of the check to the value of the information it would buy |
| D7 Judging people and organisations | Credential or tone reading only | Indicator plus base rate | Indicator, base rate, and the specific evidence that would overturn the read |

### Step 2

Aggregation and calibration are computed together, because the calibration number is only interpretable alongside its coverage. A Brier score computed over three of twenty claims is a weak measurement, not a clean one.

```
gates = input_state["gates"]      # {gate_id: bool}
dims = input_state["dims"]        # {dim_id: (score_0_4, weight)}
claims = input_state["claims"]    # list of {stated_p, outcome_or_None}

passed = all(gates.values())
num = sum(s * w for s, w in dims.values())
den = sum(w for _, w in dims.values()) or 1.0
raw = num / den
```

```
9
10 # a failed gate caps the overall result below the comfortable range
11 cap = 2.5
12 overall = raw if passed else min(raw, cap)
13
14 resolved = [c for c in claims if c["outcome"] is not None]
15 coverage = len(resolved) / len(claims) if claims else 0.0
16
17 if resolved:
18     brier = sum((c["stated_p"] - c["outcome"]) ** 2 for c in
      resolved) / len(resolved)
19     mean_p = sum(c["stated_p"] for c in resolved) / len(resolved)
20     mean_o = sum(c["outcome"] for c in resolved) / len(resolved)
21     probability_bias = mean_p - mean_o
22 else:
23     brier = None
24     probability_bias = None
25
26 output_state["overall_score"] = overall
27 output_state["failed_gates"] = [g for g, ok in gates.items() if not
      ok]
28 output_state["brier"] = brier
29 output_state["probability_bias"] = probability_bias
30 output_state["calibration_coverage"] = coverage
31 output_state["n_claims"] = len(claims)
```

Interpretation rules. The signed probability bias is the mean stated probability minus the observed event frequency among resolved claims. It can hide opposing errors and does not measure overall calibration or confidence extremity. Both it and the Brier score are limited by small samples and selection into resolvable claims, so report coverage and sample size. Write scores before consulting a model answer and cite transcript evidence for contested judgments. A failed integrity gate remains a different diagnosis from weak performance with all gates passed.

- Voice merging: attributing an individual staff view, such as a per-person giving suggestion or an informal disagreement post, to the institution.
- Dated as current: citing an old institutional timeline estimate as if it were a present forecast.
- Comparator as neutral: treating a fund that argues for its own model as a disinterested description of that model.
- Numeric laundering: repeating a source's qualitative hedge as though it were a probability, or the reverse.
- Crux as topic: naming an area of disagreement instead of the premise whose flip changes the decision.

- Undenominated benefit: reporting a benefit figure with no denominator or baseline, which the earlier uncertainty-register work already flagged as a recurring defect.

## Sub-Question 4: Subquestion 4: Triage, revise, and run the second pass on fresh prompts

Rank revisions by expected marginal score gain per minute rather than by lowest score, repair gate failures first, and then run the second pass on new prompts of the same type so that the improvement is measured as transfer rather than as memorisation. Close with an honest list of what the loop cannot repair.

### Step 1

The first non-trivial move in triage is refusing to rank revisions by lowest score. Fixing a dimension that sits at 2 out of 4 is worth little if it is not decisive, while a gate failure caps the entire result no matter how polished the rest is. Revisions are therefore ranked by expected marginal gain, and the expected gain of a gate repair is not additive but multiplicative, because passing the gate lifts the cap on everything else.

$$P_r = \frac{\Delta\mathbb{E}[S \mid r]}{m_r}, \qquad \Delta\mathbb{E}[S \mid \text{gate}] \gg \Delta\mathbb{E}[S \mid \text{dimension}]$$

Table 153: Triage list. Gate repairs come first because they lift the cap rather than adding a point.

| Rank | Revision | Why it buys the most per minute | Gate or dimension |
|---|---|---|---|
| 1 | Repair any failed integrity gate by relabelling the borrowing as a hypothetical and naming the true producer | A failed gate caps the whole result, so this is the only revision that can change the ceiling | Gate G1 or G2 |

| Rank | Revision | Why it buys the most per minute | Gate or dimension |
|---|---|---|---|
| 2 | Add a supported uncertainty statement; quantify it where the exercise requires it and the number is meaningful. | It converts an unscoreable answer into a scoreable one and unlocks the calibration measurement | Gate G3 |
| 3 | Replace the named topic with the named variable and its flip direction | It changes what the answer asserts rather than how it is phrased, and it is a one-line edit per answer | Gate G4 |
| 4 | Rewrite the dilution answer so that it commits to a recommendation within the budget | Decisiveness is separately scorable and cheap to fix once the underlying position is already held | |
| 5 | Attach an explicit stop rule and a cost to the diligence answer | Adding the stop rule is a few words and it is the discriminating feature between decisive and rash | |
| 6 | Add the base rate to the thin-package team read | Lower expected gain per minute, because the base rate must be sourced rather than asserted | |

- Revise the mechanism, not the wording. If an answer failed because the underlying reasoning was absent, rewriting the sentence moves no score.
- Keep the revision list observable. Each item names the transcript line it addresses and the observable that will show it is fixed.
- Do not overfit to the specific items. Revising the exact words of answer

A does not transfer; revising the reasoning pattern does, which is why the second pass uses fresh prompts.

### Step 2

Score the second pass with the same rubric and report both scores and their difference. Improvement on fresh prompts is suggestive, but a single score difference cannot distinguish transfer from easier questions, practice effects or scoring drift. Comparable prompt difficulty and repeated or independent assessment would strengthen the inference.

```
revisions = input_state["revisions"]   # [{name, delta_expected, minutes, is_gate}]

# Editorial prioritization heuristic, not a derived expected-score gain; 2x does not guarantee gates rank first.
for r in revisions:
    mult = 2.0 if r["is_gate"] else 1.0
    r["priority"] = mult * r["delta_expected"] / max(r["minutes"], 1e-9)

ranked = sorted(revisions, key=lambda r: r["priority"], reverse=True)

pass1 = input_state.get("pass1_score")
pass2 = input_state.get("pass2_score")
transfer = None if (pass1 is None or pass2 is None) else (pass2 - pass1)

output_state["ranked_revisions"] = ranked
output_state["transfer_delta"] = transfer
output_state["gate_repairs_first"] = [r["name"] for r in ranked if r["is_gate"]]
```

Limitations. A self-scored mock cannot independently validate the scorer's judgments. Brier score and signed probability bias require credible resolved outcomes; self-ratings or unresolved claims do not supply calibration ground truth. The real interview format is not established by the corpus. Some criteria, including judgment of people and organizations, also require evidence beyond a transcript and benefit from a knowledgeable independent reviewer.

- What this loop documents: observed responses under time pressure, adherence to integrity constraints, and performance differences on a second set of prompts. It supplies limited evidence about transfer, not a definitive measurement.
- *What it does not establish: real-world calibration against an external scorer, performance on an unseen format such as a live case interview, and any claim of prior experience, which remains prohibited.*

- Honest residual: a single second pass is one sample. A transfer delta computed from one pair of transcripts is an anecdote, not an estimate of the candidate's steady-state ability.