Conditional Optimality under Scalar Objectives: Expected Value, Candidate Sets, and Institutional Thresholds

Flyxion
Independent Researcher

November 2026

Abstract

Effective altruism ranks interventions by estimated good per unit of resource and recommends the highest-ranked option. This essay examines the formal structure of that procedure and separates four kinds of difficulty that discussions often run together: inconsistency, invalid inference, underdetermination of inputs by evidence, and disagreement over ethical axioms. Only the middle two are treated as mathematical results, and the assumptions each requires are stated. Eight lines of argument are developed, supported by nineteen formal results: sixteen propositions, one theorem, and two corollaries. First, with unbounded benefits and any positive probability admissible, an additive expected-value rule lets a large enough hypothetical benefit outweigh any finite certain harm. Second, a sum of individual welfare representations is not determined without interpersonal comparability assumptions. Third, the maximum over a restricted candidate set is certified as globally optimal only under a bound on excluded alternatives, and enlarging the set can raise the upward bias of the winning estimate. Fourth, an intervention has a context-independent marginal contribution only under additive separability. Fifth, observed success under an existing allocation need not identify the causal ranking of interventions. Sixth, uncertainty about the location of the optimum can be summarized by a perplexity, reported with the probability that the optimum was never evaluated, which depends on the prior and not on the evidence alone. Seventh, an institution that optimizes a measured objective can degrade the verification on which the measure depends, and a maximum confers no permission to act. Eighth, the same structure arises when a scalar predictive loss is used as a threshold for coordinating decisions about frontier AI development: average loss does not determine capability, and a threshold rule detects every hazardous model only under a bound on capability among models above it. The essay concludes that expected good, taken alone, underdetermines a legitimate decision procedure, and that bounded utility, robust criteria, side constraints, and verification requirements each relocate normative weight rather than eliminate it.

Keywords: effective altruism; expected value; candidate sets; interpersonal comparability; optimizer's curse; cluelessness; institutional verification; perplexity; capability thresholds.

1 Introduction

Effective altruism is usually described as the use of evidence and reasoning to determine how to benefit others as much as possible, and in its most quantitative form it operates by estimating the good done per unit of resource across a set of candidate interventions and directing resources to those at the top of the ranking. The ranking is the product of a computation, and a computation can be examined for the assumptions it requires independently of the people who perform it. That is the purpose of this essay. It does not argue that effective altruism is incoherent, and it does not treat disagreement with utilitarianism as a mathematical refutation, since a utilitarian may adopt any of the substantive assumptions discussed below as ethical premises. The question is narrower and, for that reason, more tractable: which assumptions must hold for the ranking to support the practical conclusion drawn from it, and what happens to the conclusion when they do not.

Four kinds of difficulty are conflated in much of the debate, and separating them is a precondition for saying anything precise. An inconsistency is a derivation of a contradiction from the stated principles. An invalid inference is a step from a correct calculation to a conclusion that the calculation does not establish. An underdetermined model is one in which the quantities that enter the calculation are not fixed by the available evidence, so that distinct choices compatible with the evidence yield distinct recommendations. A contested axiom is an ethical premise that some accept and others reject, where the disagreement is not about mathematics. The results below are of the second and third kinds. No inconsistency is claimed. The contested axioms are identified so that they can be argued on their merits and not smuggled in as consequences of arithmetic.

The essay also declines to use the conduct of any individual as evidence about the formal structure. The documented facts about the collapse of the exchange FTX are that its founder was convicted of fraud and conspiracy offences and was sentenced in March 2024 to twenty-five years in prison (United States Department of Justice, 2024). Those facts establish nothing about what reasoning, if any, motivated the offences, and nothing in the argument depends on that question. The case is useful for a different reason, which is taken up in Section 10: it makes vivid the difference between an objective that ranks actions and a decision procedure that determines which actions an agent may take with resources that belong to others.

The argument proceeds as follows. Section 2 places it among existing literatures. Section 3 treats the inputs to expected-value calculations. Section 4 treats aggregation across persons. Section 5 treats the construction of the candidate set, which is the part of the problem that the optimization does not itself perform. Section 6 shows why the value of an intervention is not in general a property of the intervention alone, and Section 7 shows that observed success under an existing allocation need not identify the causal ranking of interventions. Section 8 examines comparisons of cost and feasibility, and Section 9 examines the cost of the search for better interventions. Section 10 turns to institutions, Section 11 applies the analysis to a scalar predictive loss proposed as a threshold for coordinating decisions in frontier AI development, and Section 12 asks what the available repairs cost. Section 13 states what has and has not been shown.

2 Relation to Existing Work

Several literatures bear on the problems discussed here, and the contribution of this essay is to connect them through one question, namely which parts of a cost-effectiveness ranking are determined by the evidence and which are determined by the framing of the decision problem. The problem of cluelessness, introduced for consequentialism by Lenman (2000) and developed for effective altruism by Greaves (2016) and Mogensen (2021), concerns the difficulty of identifying even the sign of long-run consequences. Herlitz (2019) argues that this difficulty leaves effective altruist principles practically incomplete and proposes a decision method that works with such incompleteness. The present essay does not depend on any resolution of that debate; it asks instead what follows for rankings when the candidate set and the interactions among interventions are themselves uncertain.

Côté and Steuwer (2023) argue that charity assessment methods presuppose that the benefit of an intervention varies continuously with its scale, and that this presupposition fails in a wide class of cases, with consequences for how rankings should be read. Their analysis is a direct precursor of Section 8. Beckstead and Thomas (2024) show that every theory of the value of uncertain prospects must be reckless about tiny probabilities, timid about them, or non-transitive, and Wilkinson (2022) defends the fanatical horn of that trilemma; Section 3 uses a much weaker observation, that an additive rule with unbounded benefits cannot by itself exclude the conclusion. Greaves et al. (2024) show that on some natural ways of making a desire to make a difference precise, such preferences diverge from the desire that the world be better, a point that connects with the treatment of marginal contribution in Section 6. The selection effect that arises when the best of several noisy estimates is chosen is known as the optimizer's curse (Smith and Winkler, 2006), and Section 5 shows how it interacts with the problem of omitted alternatives. The relation between a small decision problem and the grand problem of which it is a restriction goes back to Savage (1954), who treated the choice of a small world as a modelling decision outside the theory. The identification of causal effects from observational data is treated in the potential-outcomes and graphical traditions (Pearl, 2009) and through partial identification (Manski, 1990); Section 7 applies that established theory to rankings of interventions and claims no new theorem. The case in Section 11 concerns a proposal by Pachocki (2026), and the cross-entropy decomposition used there is standard (Cover and Thomas, 2006).

3 Expected Value and Its Inputs

The basic expression is

$$V(a) = \sum_{i=1}^{n} p_i(a)\, U_i(a), \tag{1}$$

where a is an action, $p_i(a)$ is the probability that outcome i follows from it, and $U_i(a)$ is the moral value of that outcome. The expression is mathematically unobjectionable when the probabilities and values are well defined. Difficulties arise when the probabilities rest on speculative causal

models and not on repeated observation, when the values combine consequences of different kinds, and when the values are allowed to be arbitrarily large. Consider an action with a one-in-a-million probability of a benefit equivalent to saving a billion lives. Then $V = 10^{-6} \times 10^{9} = 10^{3}$, which is an expected benefit of one thousand lives. The multiplication is correct. What the multiplication cannot supply is a reason to believe that the probability is 10^{-6} and not 10^{-9} or zero, and the inference from a large product to a moral preference for the action is not licensed by the product alone.

The structural point can be stated precisely. Let a_0 be a status quo with $V(a_0) = 0$, and let $a(H, p, B)$ denote an action that imposes a certain disvalue $H > 0$ and delivers a benefit B with probability $p > 0$.

Proposition 1 (Speculative-benefit domination)**.** *Suppose the value scale is unbounded above, every positive probability is admissible, and actions are ranked by the additive expected value* (1)*. Then for every finite $H > 0$, every $p \in (0, 1]$, and every finite value v, there exists B^* such that $V(a(H, p, B)) > v$ whenever $B > B^*$.*

Proof. For this action $V = pB - H$, which exceeds v exactly when $B > (H + v)/p$. Take $B^* = (H + v)/p$. □

The proposition is elementary, and it is stated because its consequences are easily misdescribed in both directions. It does not show that proponents of expected-value reasoning endorse harmful conduct, and it does not show that large benefits are not sometimes worth pursuing. It shows that the additive rule with unbounded benefits does not, on its own, contain any protection against the conclusion that a sufficiently large hypothetical benefit justifies a certain harm. Any such protection must come from elsewhere: from a bound on B, from a lower bound on admissible probabilities, from a restriction on which harms may be imposed on whom, or from a requirement that the benefit be verifiable. Each of these is an additional commitment, and the choice among them changes the recommendation.

The role of the probability deserves separate emphasis because the product pB is the only quantity that enters the calculation. A halving of p is exactly offset by a doubling of B, so an estimate that places a speculative benefit at an astronomical scale will dominate unless the probability is pushed below the reciprocal of that scale, and the evidence available for probabilities of that order is almost never of a kind that could justify them. If credence is represented not by a single number but by a set $[p_{\mathrm{lo}}, p_{\mathrm{hi}}]$, the worst-case value of the action is $p_{\mathrm{lo}}B - H$, and when $p_{\mathrm{lo}} = 0$ this equals $-H < 0$ for every B, so the domination result disappears under a maximin criterion (Gilboa and Schmeidler, 1989) while remaining intact under any single precise prior. The difference between the two verdicts is not a difference in arithmetic. It is a difference in whether the evidence is taken to fix a point probability, which is the distinction between risk and ambiguity drawn by Ellsberg (1961).

The qualifier that every positive probability is admissible carries much of the weight and should be separated into three sets. Let $\mathcal{P}_{\mathrm{logical}}$ denote the probability assignments that are logically possible, $\mathcal{P}_{\mathrm{model}}$ those permitted by a chosen model, and $\mathcal{P}_{\mathrm{evidence}}$ those compatible with the available evidence. Proposition 1 requires only the first. The failure that matters in

practice occurs when a calculation substitutes a probability drawn from the first or second set for one justified by the third, and its force depends on whether $\mathcal{P}_{\text{evidence}}$ excludes the small values on which the domination relies. This is not an argument that tiny probabilities are illegitimate, since evidence can sometimes support them; it is an argument that the admissible set must be justified and not assumed.

A parallel distinction applies to the magnitude of the benefit. A benefit may be logically conceivable, physically feasible, or evidentially supported, and Proposition 1 requires only an unbounded domain of the first kind. The defect therefore lies in two places, the probability assignment and the benefit domain, and repairing the first while leaving the second unrestricted does not remove the dependence of the conclusion on how large a benefit the model is allowed to contain.

The astronomical numbers that appear in arguments about the far future arise from a particular source, which is the product of an assumed population size and an assumed per-person welfare, as in the argument of Bostrom (2003). The population enters as a parameter N in a total of the form $B = N\bar{w}$, and Proposition 1 applies to any such total for sufficiently large N. Whether the total view of welfare, which counts merely possible persons by the same additive rule, is correct is an ethical question on which the mathematics is silent; the results of Arrhenius (2000) show that the axioms one might wish to hold of an additive population ethics are jointly unsatisfiable, so that some premise must be given up, and the premise given up determines how large B can become. The relevant observation here is only that the size of B is itself a modelling choice.

4 Aggregation Across Persons

The social value of an outcome is commonly written

$$W = \sum_{j=1}^{N} u_j, \tag{2}$$

where u_j represents the welfare of person j. For (2) to have a quantitative interpretation, there must be a fact about how a unit of welfare for person j compares with a unit for person k. Ordinary expected-utility theory does not supply this fact. Each person's preferences over lotteries determine a utility function only up to a positive affine transformation,

$$u_j' = \alpha_j u_j + \beta_j, \qquad \alpha_j > 0, \tag{3}$$

and the transformation may be chosen independently for each person without altering any person's preferences.

Proposition 2 (Representation dependence of the unweighted sum)**.** *Let x and y be outcomes such that person 1 strictly prefers x and person 2 strictly prefers y. Then the sign of $W(x) - W(y)$ computed from* (2) *is not invariant under transformations of the form* (3)*.*

Proof. Additive constants β_j cancel in the difference between outcomes, so only scales matter. We have $W(x) - W(y) = (u_1(x) - u_1(y)) + \alpha_2(u_2(x) - u_2(y))$. The first term is positive and the coefficient of α_2 is negative, so as α_2 ranges over $(0, \infty)$ the expression changes sign. □

A concrete case makes the point plain. Let $u_1(x) = 2$, $u_1(y) = 0$, $u_2(x) = 0$, $u_2(y) = 1$. Then $W(x) = 2 > W(y) = 1$. Replacing u_2 by $3u_2$, which represents exactly the same preferences, gives $W(x) = 2 < W(y) = 3$. The ranking of the social outcomes has reversed although nothing about any individual's preferences has changed.

This is not a refutation of utilitarianism. A utilitarian may hold that there is a fact of the matter about interpersonal comparisons of welfare, and the social choice literature has characterized the conditions under which the utilitarian sum is the unique rule consistent with comparability assumptions of a stated kind (Sen, 1970; d'Aspremont and Gevers, 1977). Harsanyi (1955) derives the utilitarian form from axioms that already include interpersonal comparability as an input. The point is that probability theory and individual rationality do not establish those comparisons, so the numerical value of any large aggregate depends on a premise that must be defended on ethical and empirical grounds, and a cost-effectiveness figure expressed in a common unit, such as units of welfare per dollar, inherits that dependence. The dependence is amplified when the comparison spans species, or spans persons who do not yet exist, because in those cases there is no direct preference evidence from which a scale could even be estimated.

5 The Candidate Set

An optimization procedure returns the best element of a set that someone has specified. The specification of that set is not itself an output of the optimization, and the conclusion that an intervention is the most effective way of doing good is a claim about the whole space of possible interventions, whereas any evaluation covers a subset. Let Ω denote the space of feasible interventions and portfolios of interventions (restricted, for the purposes of this section, to those that are also permitted, since Section 10 separates feasibility from permission), let $S \subset \Omega$ denote the alternatives actually evaluated, let V denote the true value, and let $\widehat{V}$ denote the evaluator's estimate. The evaluator computes

$$a^*_S = \arg\max_{a \in S} \widehat{V}(a), \tag{4}$$

whereas the quantity of practical interest is

$$a^*_\Omega = \arg\max_{a \in \Omega} V(a). \tag{5}$$

Proposition 3 (Uncertified counterfactual selection)**.** *Let S be a proper subset of Ω. If no upper bound is available on the value of the alternatives in $\Omega \setminus S$, then identifying the maximizer over S provides no guarantee that it is optimal over Ω, even when the values on S are known exactly.*

Proof. Fix any assignment of values on S and let m be its maximum. Choose an alternative $b \in \Omega \setminus S$ and assign it any value greater than m. Every comparison within S is unchanged, and $a^*_S \neq a^*_\Omega$. □

The proof is immediate, and the statement is still worth making because it identifies exactly what a claim of global optimality requires. The following result states the requirement in general form and allows the evidence to leave the values on S uncertain as well. Let $\mathcal{V}$ be the set of value functions on Ω compatible with the evidence, and call a candidate a° robustly optimal if $V(a^\circ) \geq V(a)$ for every $a \in \Omega$ and every $V \in \mathcal{V}$.

Theorem 1 (Certification condition)**.** *A candidate $a^\circ \in S$ is robustly optimal if and only if*

$$\sup_{V \in \mathcal{V}} \sup_{a \in \Omega} \left[V(a) - V(a^\circ) \right] \;\leq\; 0. \tag{6}$$

The left side equals $\max\{\rho, \omega\}$, where $\rho = \sup_V \sup_{a \in S}[V(a) - V(a^\circ)]$ is the ranking term and $\omega = \sup_V \sup_{a \in \Omega \setminus S}[V(a) - V(a^\circ)]$ is the omission term, with the convention that a supremum over an empty set is $-\infty$. Since the choice $a = a^\circ$ contributes zero, $\rho \geq 0$, and the condition is equivalent to $\rho = 0$ and $\omega \leq 0$. If a° maximizes every $V \in \mathcal{V}$ over S, which is stronger than maximizing a single estimated value function, then $\rho = 0$ and the left side is $\max\{0, \omega\}$, so the condition reduces to $\omega \leq 0$. When the evidence fixes the values on S this reads $\sup_{V \in \mathcal{V}} \sup_{a \in \Omega \setminus S} V(a) \leq m$, where m is the known maximum over S. In particular, if $\mathcal{V}$ places no upper bound on $V(b)$ for some excluded b, then $\omega = +\infty$ and no element of S is certified.

Proof. Robust optimality means that $V(a) - V(a^\circ) \leq 0$ for all a and V, which is (6) after taking suprema. Since $\Omega = S \cup (\Omega \setminus S)$, the supremum over Ω is the larger of ρ and ω. Taking $a = a^\circ$ shows $\rho \geq 0$, so (6) holds if and only if $\rho = 0$ and $\omega \leq 0$. If a° maximizes every V over S, every term in ρ is at most zero, so $\rho = 0$. If some excluded b admits arbitrarily large $V(b)$, then $\omega = +\infty$. □

The theorem also supports an approximate version. Define the worst-case regret $\mathcal{R}(a^\circ) = \sup_{V \in \mathcal{V}} \sup_{a \in \Omega}[V(a) - V(a^\circ)]$, and call a° ε-robustly optimal if $\mathcal{R}(a^\circ) \leq \varepsilon$.

Corollary 1 (Approximate certification)**.** *$\mathcal{R}(a^\circ) = \max\{\rho, \omega\}$, so for $\varepsilon \geq 0$ the candidate a° is ε-robustly optimal if and only if the ranking regret satisfies $\rho \leq \varepsilon$ and the omission regret satisfies $\omega \leq \varepsilon$. If the values on S are fixed, a° attains their maximum m, and the evaluator can assert an upper bound U on the value of every excluded alternative, then $\mathcal{R}(a^\circ) \leq \max\{0, U - m\}$.*

Proof. The first statement is the decomposition in Theorem 1 with ε in place of zero. For the second, $\rho = 0$ because the values on S are fixed and a° attains their maximum, and $\omega \leq U - m$ by the asserted bound. □

An evaluator who cannot demonstrate global optimality may therefore still establish that no compatible alternative improves welfare by more than a stated amount. The two sources of regret can be bounded separately, the first by sensitivity analysis and the second only by an assertion about the search, such as the bound U.

The theorem distinguishes two kinds of uncertainty that are often reported together. Ranking uncertainty is uncertainty about the order of the elements of S, and it is the kind addressed by conventional sensitivity analysis. Omission uncertainty is uncertainty about the term ω in (6), and it cannot be addressed by varying the parameters of the models for the elements of S, because the omitted alternatives are not in those models. A defensible claim that follows from a careful evaluation is that an intervention performs best among those evaluated under a stated model, which may be called conditional optimality. The stronger claim of unqualified optimality requires the bound in (6), and a bound of that kind is an assertion about the search that produced S and about the unobserved opportunity space.

The two kinds of uncertainty can be summarized in a single quantity by borrowing the notion of perplexity, which is the exponential of entropy and is read as the effective number of equally likely alternatives (Jelinek et al., 1977). Let μ be a prior over the evidence-compatible set $\mathcal{V}$, and assume for simplicity that each $V \in \mathcal{V}$ has a unique maximizer over Ω. This assumption matters for part (ii) below, since robust optimality as defined permits a candidate to share the maximum with another; with ties, π would require a tie-breaking rule. For $a \in S$ let $\pi(a) = \mu(\{V : \arg\max_\Omega V = a\})$, and let $\pi_0 = \mu(\{V : \arg\max_\Omega V \in \Omega \setminus S\})$ be the mass assigned to the possibility that the best intervention was never evaluated. Define

$$H(\pi) = -\sum_{a \in S} \pi(a) \ln \pi(a) - \pi_0 \ln \pi_0, \qquad \mathrm{PP}(\pi) = \exp H(\pi). \tag{7}$$

Grouping the omitted alternatives into one cell can only lower the entropy, so PP as defined is a lower bound on the perplexity over all of Ω.

Proposition 4 (Perplexity of the location of the optimum)**.** *(i)* $1 \le \mathrm{PP}(\pi) \le |S| + 1$*, with* $\mathrm{PP}(\pi) = 1$ *if and only if all of the mass lies in a single cell. (ii) A candidate* $a^\circ \in S$ *is robustly optimal if and only if* $\pi(a^\circ) = 1$ *for every prior* μ *on* $\mathcal{V}$*, and in that case* $\mathrm{PP} = 1$ *for every prior. (iii) If* $\pi_0 > 0$ *and* $\pi(a) > 0$ *for some* $a \in S$*, then* $\mathrm{PP}(\pi) > 1$*. (iv) If* $\mathcal{V}$ *contains two functions whose maximizers lie in different cells, then there are priors on* $\mathcal{V}$ *with* $\mathrm{PP} = 1$ *and priors with* $\mathrm{PP} = 2$*; if both maximizers lie in the omitted cell, every mixture of the two has* $\mathrm{PP} = 1$*. (v)* $\mathrm{PP} = 1$ *does not identify the winner and is compatible with* $\pi_0 = 1$*.*

Proof. Part (i) is the standard range of the exponential of Shannon entropy over a distribution with $|S| + 1$ cells, with the minimum attained exactly at point masses. For (ii), if a° is robustly optimal then every $V \in \mathcal{V}$ is maximized at a°, so $\pi(a^\circ) = 1$ for every μ. Conversely, take μ to be a point mass at an arbitrary V; then $\pi(a^\circ) = 1$ says that V is maximized at a°, and V was arbitrary. Part (iii) follows from (i), since two cells carry positive mass. For (iv), a point mass at either function gives $\mathrm{PP} = 1$; if the maximizers lie in different cells, the equal mixture places mass one half on each of two cells, so $H = \ln 2$ and $\mathrm{PP} = 2$, whereas if both lie in the omitted cell every mixture places all its mass on that one cell. For (v), point masses on distinct cells give $\mathrm{PP} = 1$ with different winners, and a point mass on the omitted cell gives $\mathrm{PP} = 1$ with $\pi_0 = 1$. □

Proposition 5 (Dependence on the granularity of the partition)**.** *Let a cell* c *carry mass* $\pi_c > 0$ *and be refined into* m *subcells among which the prior divides that mass equally, as it would if*

the subcells were equivalent variants of one intervention. Then the entropy increases by $\pi_c \ln m$ *and the perplexity is multiplied by* m^{π_c}*. In particular, if all of the mass lay in one cell, the perplexity rises from* 1 *to* m *although no substantive uncertainty has been added.*

Proof. By the grouping property of entropy, the refined entropy equals $H(\pi)$ plus π_c times the entropy of the uniform distribution on m cells, which is $\pi_c \ln m$. Exponentiating gives the factor m^{π_c}. □

Candidate-location perplexity is therefore comparable across evaluations only when the cells are defined by a common taxonomy, and the entropy should be computed over equivalence classes of interventions and not over their arbitrary descriptions. The omitted cell is itself a coarse grouping of this kind.

Perplexity is useful here in two ways, and it should not be reported alone. It replaces the margin between first and second place, which says little when several candidates are close, with the effective number of contenders across the represented cells. It does not say where the optimum lies or whether it was evaluated, since distributions concentrated on different cells, including the omitted one, all have perplexity one. The informative report is therefore the pair $(\mathrm{PP}(\pi), \pi_0)$, the first measuring dispersion across the represented cells and the second the probability that the optimum lies outside the evaluated set. Naming a single winner need not assert certainty, since an evaluator may rationally select the most probable winner while acknowledging substantial uncertainty, and what the pair makes visible is how much. Theorem 1 gives the condition under which the winner is optimal for every prior. Part (iv) states the limit of the device: perplexity summarizes an assumed distribution and cannot justify it, and the argument of this section is that the distribution over omitted possibilities is not in general identifiable from the evidence. The measure reports the dependence of the ranking on the framing and does not remove it.

Omission uncertainty also has two components that the single mass π_0 combines. One is existence, whether any excluded alternative exceeds the best evaluated one, and the other is magnitude and identity, how much better it is and which it is. An evaluator may know that a superior intervention exists without knowing what it is, or may know that several exist without knowing whether any is superior. The expected shortfall $\mathbb{E}\big[\big(\sup_{a\in\Omega\setminus S} V(a) - V(a^\circ)\big)^+\big]$ captures magnitude where π_0 captures only probability, and the interpretation of either depends on a prior over the whole opportunity space. The broader point is that the uncertainty of a decision is not exhausted by uncertainty about the values of the alternatives that the decision-maker has chosen to compare.

Perplexity should not be read as a measure of decision quality. It counts the effective locations of the optimum and is silent on the consequences of choosing wrongly: two equally likely interventions of nearly identical value and two equally likely interventions whose values differ greatly both give perplexity two. The expected regret $\mathbb{E}[V(a^*_\Omega) - V(a^\circ)]$ and the expected shortfall measure the latter and are more directly relevant to decision quality. Together with Proposition 5, this shows that perplexity is a property of a probability distribution together with a representation of the alternatives, and not an intrinsic property of the decision problem.

A further question is whether a recommendation survives the discovery of additional alternatives, which is an operational criterion short of robust optimality over all of Ω. Let $S_1 \subset S_2 \subset \Omega$, with values known, and let $a^*(S_1)$ maximize V over S_1.

Proposition 6 (Stability under expansion)**.** *The winner $a^*(S_1)$ remains a maximizer over S_2 if and only if $\sup_{a \in S_2 \setminus S_1} V(a) \leq V(a^*(S_1))$. If the additional alternatives are k independent draws of value from a continuous distribution F, and $m = V(a^*(S_1))$, the probability that the recommendation survives all k is $F(m)^k$.*

Proof. The first statement is the definition of the maximum over $S_2 = S_1 \cup (S_2 \setminus S_1)$. For the second, the recommendation survives exactly when every one of the k independent draws is at most m, which has probability $F(m)^k$. □

The survival probability depends on F, which describes the unobserved opportunity space and is not identified by the evidence, so it expresses the stability of a recommendation under an assumed search process and not its robust optimality. It nevertheless gives a principled reading of when further evaluation is worth its cost: the expected improvement from k more evaluations is weighed against their cost, and both depend on the tail of F.

Consider four evaluators. The first has examined one hundred charities and selected the best. The second has examined ten thousand and selected a different one. The third has also considered public infrastructure, educational reform, scientific research, and changes in law. The fourth has additionally counted the opportunity cost of the evaluators' own labour, the social consequences of concentrating philanthropic influence, and the possibility that different economic arrangements would remove the need for several of the charities. Each may be perfectly consistent within the set that was compared, and yet they solve different problems. The selected action is a function $a^*(S, M, U, B)$ of the candidate set S, the causal model M, the welfare function U, and the resource constraint B, and there is no reason to expect it to be invariant as these inputs change.

The enlargement of the candidate set, which is the natural response to omission uncertainty, has a cost that is easily overlooked. When the best of several noisy estimates is selected, the winning estimate is biased upward, a phenomenon called the optimizer's curse (Smith and Winkler, 2006).

Proposition 7 (Selection bias)**.** *Let $\widehat{V}(a) = V(a) + \varepsilon_a$ with $\mathbb{E}[\varepsilon_a] = 0$ for every a in a finite set S, and let $\hat{a} = \arg\max_{a \in S} \widehat{V}(a)$. Then*

$$\mathbb{E}[\widehat{V}(\hat{a})] \;\geq\; \max_{a \in S} V(a) \;\geq\; \mathbb{E}[V(\hat{a})]. \tag{8}$$

Proof. Let a_0 maximize V over S. Since $\widehat{V}(\hat{a}) \geq \widehat{V}(a_0)$ pointwise, taking expectations gives $\mathbb{E}[\widehat{V}(\hat{a})] \geq V(a_0)$. Since $V(\hat{a}) \leq V(a_0)$ pointwise, $\mathbb{E}[V(\hat{a})] \leq V(a_0)$. □

The winner's estimated value therefore overstates its true expected value, in the sense of the displayed inequality. The size of the overstatement depends on the number of candidates

and on the error distribution, and it is not monotone in the candidate set in general, because a larger set may also contain a better candidate whose selection raises the true value of the winner. A special case shows the mechanism most clearly. For n options of equal true value and independent normal errors with standard deviation σ, the expected maximum of the estimates exceeds the common true value by an amount of order $\sigma\sqrt{2\ln n}$ for large n, so in that case the overstatement grows with n. The consequence is a statistical cost of enlarging the search and not a universal monotonicity. Enlarging S can reduce the omission term ω in (6) while exposing the evaluation to a larger selection bias, unless the estimates are shrunk toward a prior. The appropriate shrinkage depends on the distribution of true values across the alternatives, which is the feature of the opportunity space that the evidence does not reveal. The two problems therefore cannot be treated as independent, since each repair requires an assumption about what the other is trying to learn.

The trade-off can be written as a loss $\mathcal{L}(S) = \mathbb{E}[V(a^*_\Omega) - V(a^*_S)] + C(S)$, where a^*_S is selected using noisy estimates and $C(S)$ includes the cost of discovering, evaluating, and verifying the candidates. Under exact values and zero cost, enlarging S weakly improves the optimum. Under noisy selection that monotonicity can fail, and an example shows how severely.

Proposition 8 (Non-monotonicity of realized selection)**.** *Let $S_1 = \{a\}$ with $V(a) = 1$, and let $S_2 = S_1 \cup \{c\} \cup \{b_1, \ldots, b_n\}$ with $V(c) = 1 + \Delta$ for some $\Delta > 0$ and $V(b_i) = 0$. Suppose the estimates are $\widehat{V}(x) = V(x) + \varepsilon_x$ with independent $N(0, \sigma^2)$ errors and $\sigma > 0$. Then the selected intervention from S_1 has expected true value 1, while the selected intervention from S_2 has expected true value tending to 0 as $n \to \infty$, although $\max_{S_2} V = 1 + \Delta > 1$.*

Proof. From S_1 the only candidate is a, with value 1. For S_2, let E_n be the event that $\max_i \varepsilon_{b_i} > \max\{1 + \varepsilon_a,\, 1 + \Delta + \varepsilon_c\}$. When E_n occurs, some b_i has the highest estimate and the selected value is 0. The maximum of n independent normal errors diverges almost surely as $n \to \infty$ while the right side is a fixed finite random variable, so $P(E_n) \to 1$. The selected true value never exceeds $1 + \Delta$, so its expectation is at most $(1 + \Delta)P(E_n^c) \to 0$. □

Expanding the candidate set to include a strictly better alternative can therefore lower the expected true value of the selected intervention below that obtained from the smaller set, and with positive evaluation costs the loss increases further. The result connects the statistical problem to the institutional decision about how extensively to search.

6 Interaction and the Failure of Separability

The preceding sections treated each intervention as having a value of its own. Suppose that a fixed budget will educate some set S of children, and let $W(S)$ denote the value of the resulting world. The marginal value of adding child i to the set is

$$\Delta_i(S) = W(S \cup \{i\}) - W(S). \tag{9}$$

Whether $\Delta_i(S)$ depends on S determines whether an intervention can be assigned a value independently of the others.

Proposition 9 (Separability)**.** *$\Delta_i(S)$ is independent of S for every i and every S not containing i if and only if $W(S) = W(\emptyset) + \sum_{i \in S} w_i$ for some constants w_i.*

Proof. If W has the stated form, then $\Delta_i(S) = w_i$ for all S. Conversely, if $\Delta_i(S) = \Delta_i(\emptyset) =: w_i$ for all S, then adding the elements of S one at a time gives $W(S) = W(\emptyset) + \sum_{i \in S} w_i$. □

An evaluator can of course rank individual interventions by their marginal effects at a specified baseline without assuming global separability, and portfolio optimization can accommodate complementarity and substitution, although often at considerable computational cost. The difficulty arises when individual rankings are used to construct a portfolio without recalculating marginal effects as the portfolio changes. A fixed ranking of interventions by independently estimated value suffices for portfolio selection only under additional structural assumptions. Additive separability is one such condition, and weaker conditions can sometimes support approximation guarantees, but without appropriate structure individual rankings need not identify even an approximately optimal portfolio. A small example shows how severe the failure can be.

Example 1 (Complementarity)**.** Let there be three interventions with a budget that permits two. Let $W(\emptyset) = 0$, $W(\{1\}) = W(\{2\}) = 0$, $W(\{3\}) = \varepsilon$ for some $0 < \varepsilon < 1$, $W(\{1,2\}) = 1$, and $W(\{1,3\}) = W(\{2,3\}) = \varepsilon$. Evaluated one at a time, intervention 3 has the highest value and interventions 1 and 2 have none, so a procedure that ranks individual interventions selects intervention 3 and then one of the others, for a value of ε. The best pair is $\{1,2\}$ with value 1, and neither of its members would have been selected. The ratio of the selected portfolio's value to the optimum is ε, which can be made arbitrarily small, so no approximation guarantee holds for this ranking.

The example is a stylized version of the observation that a scientist's contribution may require a teacher, a technician, and a collaborator, so that the value of educating any one of them is conditional on the others. Complementarity is one direction in which separability fails. Substitution is another. If interventions 1 and 2 each secure the same outcome of value 1 and secure it jointly no more than singly, so that $W(\{1\}) = W(\{2\}) = W(\{1,2\}) = 1$, then summing their individually estimated values counts the outcome twice. A third failure arises in the evaluation of an institution's own contribution. If a joint project requires two organizations and yields value $W(\{1,2\}) = 1$ with no value from either alone, each organization's marginal contribution, computed as in (9) against the situation in which it does not participate, is the full value 1, and the two claimed contributions sum to 2. A preference for actions with large measured marginal contribution therefore need not coincide with a preference for a better world, which is the divergence analyzed by Greaves et al. (2024) in a different setting.

The dimension of the problem should also be stated plainly. Even if each of n interventions is a binary choice, there are 2^n portfolios. For a population of one million children the count is 2^{10^6}, which is approximately 10^{301030}. Nothing here requires that the portfolios be enumerated, since structure can be exploited and approximations can be made, and the claim is not that useful partial comparisons are impossible. The claim is that the complete counterfactual space

is not a practical object of direct comparison, so any ranking depends on a decision about which interactions to model and which to neglect, and that decision is part of the conclusion and not a preliminary to it. The condition that any child might become a distinguished scientist if given the opportunity compounds the difficulty, because it implies that the distribution of potential contributions is not sharply concentrated on an identifiable subset of children, and the evidence for ranking one child's expected contribution above another's is correspondingly weak compared with the evidence about the surrounding institutions that determine whether any contribution is realized.

7 Identification of Counterfactual Values

The preceding sections treated the value of an intervention as a quantity to be estimated. A prior difficulty is that observed success under an existing allocation may not identify that value at all. Let $A \in \{0, 1\}$ indicate whether a child receives an intervention such as advanced scientific education, let $Y \in \{0, 1\}$ indicate a subsequent achievement, and let $Y(a)$ denote the potential outcome under intervention a. Observation can identify $\mathbb{E}[Y \mid A = a]$, whereas the quantity relevant to allocation is $\mathbb{E}[Y(a)]$. The two differ when selection into the intervention depends on characteristics that also affect the outcome, and children who receive advanced education may differ systematically from those who do not. This is established theory, presented here as an application to rankings of interventions and not as a new result.

Proposition 10 (Non-identification of intervention rankings)**.** *Let (A, Y) be binary with $P(A = 1) = \pi \in (0, 1)$ and $q_a = P(Y = 1 \mid A = a)$, where $q_1 > q_0$. There exist two causal models M_1 and M_2, each generating exactly this joint distribution, such that $\mathbb{E}_{M_1}[Y(1)] - \mathbb{E}_{M_1}[Y(0)] = 0$ and $\mathbb{E}_{M_2}[Y(1)] - \mathbb{E}_{M_2}[Y(0)] = q_1 - q_0$. Consequently, if an alternative intervention is known to raise the same outcome by τ with $0 < \tau < q_1 - q_0$, the two models rank the alternative and the intervention in opposite order.*

Proof. In M_2, let $A \sim \text{Bernoulli}(\pi)$ and let $Y \mid A = a \sim \text{Bernoulli}(q_a)$ with A exogenous, so that $\mathbb{E}[Y(a)] = q_a$. In M_1, let $Y \sim \text{Bernoulli}(\bar{q})$ with $\bar{q} = \pi q_1 + (1 - \pi)q_0$, and let $A \mid Y = y \sim \text{Bernoulli}(\pi_y)$ with $\pi_y = P(A = 1 \mid Y = y)$ computed from the given joint distribution by Bayes' rule, with the outcome unaffected by A, so that $Y(1) = Y(0) = Y$. Both models reproduce the joint distribution of (A, Y). The effect is zero in M_1 and $q_1 - q_0$ in M_2, which gives the stated reversal. □

The result is an instance of a general fact: the distribution of observed variables restricts the set of compatible causal models without selecting one. Without assumptions about the selection mechanism, the worst-case bounds on the average effect of a binary-outcome intervention have width one on the unit interval (Manski, 1990), and narrowing them requires identifying assumptions such as randomization, valid instruments, or ignorability, each of which is a claim about how interventions were assigned and is not tested by the observed distribution. The implication for the counterfactual space is stronger than its sheer size. Even exhaustive observation of the existing world need not identify the causal ranking of interventions in

alternative worlds, so the enlargement of the evidence base does not by itself close the gap between a ranking of observed outcomes and a ranking of what would happen under reallocation. In the vocabulary of Proposition 4, for a prior supported on these observationally equivalent models, repeated observations of the same variables do not change their relative posterior probabilities, so the perplexity of the posterior over them equals that of the prior however much data of this kind accumulates. Identification requires additional information or assumptions that distinguish their causal structures, such as interventions, further variables, or structural restrictions.

8 Cost, Feasibility, and Optimization over Claims

A comparison of the cost of achieving an outcome is a comparison of the lowest quoted costs only if the proposals are equivalent in everything else. Suppose three entrepreneurs propose a recycling facility at capital costs $c_1 < c_2 < c_3$, with respective probabilities of success p_k and benefits b_k if successful. If failure yields nothing and the capital is lost, then the expected net benefit of proposal k is $V_k = p_k b_k - c_k$, and proposal 1 is preferred to proposal 2 exactly when $p_1 b_1 - p_2 b_2 > c_1 - c_2$. This inequality shows that ranking by cost is correct only when $p_k b_k$ is common across proposals. In general the cheapest proposal is the best only if the cheapness does not reflect a lower probability of success or a smaller benefit, and the quoted capital cost is not a sufficient statistic for the intervention. It is also relevant that each of the three quantities may be misreported, that the facility may displace an existing operation, and that the benefits may be of a kind that cannot be added to the costs.

The continuity assumption identified by Côté and Steuwer (2023) deserves mention here. If the benefit $B(c)$ of spending c on a facility is zero below a threshold $\bar{c}$ and equal to $\bar{B}$ at or above it, then a marginal analysis evaluated at any $c < \bar{c}$ reports zero return, and the ranking of the facility against alternatives with smoothly increasing returns is wrong in the region where it matters. Interventions with threshold scale are the rule in construction and manufacturing, so the evaluation of a recycling factory cannot be reduced to an average or marginal figure without a model of the threshold.

A further issue arises when proposals are submitted by parties who know more about their own costs than the evaluator does. Let proposal k report a cost $\hat{c}_k = c_k - \delta_k$, where δ_k is the amount by which the reported cost falls below the actual cost, which the evaluator does not observe. The shortfall may reflect honest estimation error, missing information, optimism, or deliberate misrepresentation, and the mathematics does not distinguish among these. The evaluator selects the proposal with the lowest reported cost. The following result states that this procedure selects for favorable reporting error.

Proposition 11 (Selection on favorable reporting error)**.** *Let the pairs (c_k, δ_k) be independent and identically distributed across $k = 1, \dots, n$ with $\mathbb{E}|\delta_1| < \infty$, let $\hat{c}_k = c_k - \delta_k$ have a continuous distribution function F, and let K be the index of the proposal with the lowest reported cost. Define $m(t) = \mathbb{E}[\delta_1 \mid \hat{c}_1 = t]$. If m is nonincreasing, then $\mathbb{E}[\delta_K] \geq \mathbb{E}[\delta_1]$.*

Proof. The probability that proposal 1 has the lowest reported cost, given $\hat{c}_1 = t$, is $g(t) = (1 - F(t))^{n-1}$, which is nonincreasing. By Chebyshev's association inequality for two functions of $\hat{c}_1$ that are monotone in the same direction, $\mathbb{E}[m(\hat{c}_1)g(\hat{c}_1)] \geq \mathbb{E}[m(\hat{c}_1)]\,\mathbb{E}[g(\hat{c}_1)]$. The left side equals $\mathbb{E}[\delta_1 \mathbf{1}\{K = 1\}]$ and the right side equals $\mathbb{E}[\delta_1]/n$. By exchangeability, $\mathbb{E}[\delta_K] = n\,\mathbb{E}[\delta_1 \mathbf{1}\{K = 1\}] \geq \mathbb{E}[\delta_1]$. □

Corollary 2 (Log-concave case)**.** *If c_k and δ_k are independent with log-concave densities and $\mathbb{E}|\delta_1| < \infty$, then m is nonincreasing and $\mathbb{E}[\delta_K] \geq \mathbb{E}[\delta_1]$.*

Proof. Set $X = -\delta$ and $Y = c$, which are independent with log-concave densities, so that $\hat{c} = X + Y$. A classical monotonicity result for sums of independent log-concave variables (Efron, 1965) states that $\mathbb{E}[X \mid X + Y = t]$ is nondecreasing in t. Because $\delta = -X$, the conditional expectation $m(t) = -\mathbb{E}[X \mid X + Y = t]$ is nonincreasing in t, and the proposition applies. □

The proposition concerns selection on favorable reporting error and says nothing about intent. Nor does it say that the selected proposal's actual cost exceeds that of any other proposal; it concerns the expected reporting error of the selected proposal, which exceeds the unconditional mean. If applicants can choose their reports strategically, the evaluator's selection rule changes their incentives to report accurately, and a mechanism-design treatment in which the rule is chosen with those incentives in view is the natural extension. The proposition also formalizes the observation that when proposals are claims, a procedure that rewards projected cost-effectiveness is an optimization over claims and not over realizable outcomes. If proposals can be submitted without limit, there may always be another entrant promising the same result for half the cost. The remedy is independent verification of feasibility, and verification consumes resources of its own, which is the subject of the next section and, in institutional form, of Section 10.

9 The Cost of Searching

Searching for better interventions uses resources that could be spent on interventions already known. Suppose an evaluator holds a budget and may either spend all of it on the best known option or spend a fraction on investigating alternatives that might be more effective. The expected value of that investigation depends on the distribution of values among the undiscovered alternatives, and so the decision to search requires a prior over undiscovered opportunities. If that prior is itself to be justified, further inquiry is needed, which yields a recursion from optimizing the intervention to optimizing the search to optimizing the information about the search. Decision theory handles recursions of this kind through the value of information (Howard, 1966) and through optimal stopping, and there is no logical contradiction in the recursion.

The significance for the present argument is limited but real. The recursion must be terminated by an assumption about costs, uncertainty, and acceptable decision quality, and the point of termination is not itself a result of the optimization. An evaluator cannot establish

that the search was globally optimal by applying the vocabulary of optimization to it, since the claim would require a bound of the kind in (6) on the value of the searches that were not conducted. This does not make search irrational. It makes the stopping rule a premise of the recommendation.

10 Institutions, Verification, and Authority

The preceding results concern the calculation. This section concerns the setting in which the calculation is used, which is typically an organization acting on behalf of others and controlling resources that belong in part to them.

The first point is a logical one. An optimization procedure ranks actions with respect to an objective. Let

$$a^* = \arg\max_{a \in A} V(a). \tag{10}$$

This establishes that a^* maximizes the stipulated objective over the stipulated action set. It does not establish Permitted(a^*). Permission is a normative relation determined by rights, obligations, consent, fiduciary duties, and legitimate authority, and it is not a function of V. If $P \subset A$ denotes the set of permitted actions, the correct recommendation is $\arg\max_{a \in A \cap P} V(a)$, and $a^* \notin P$ implies that a^* is impermissible however large the margin $V(a^*) - \max_{a \in A \cap P} V(a)$ may be. The relevant question about any case in which an agent gambled with assets held for others is therefore whether the permission constraint was part of the objective, was imposed on it from outside, or was left to the discretion of the optimizing agent. The documented offences in the FTX case (United States Department of Justice, 2024) are consistent with any of these structures, and this is the sense in which the case is a test of the question and not a proof of an answer.

This also refines the candidate-set analysis. Let Ω_{feas} be the physically feasible universe, $P \subset \Omega_{\text{feas}}$ the permissible alternatives, and S the evaluated subset. The certification theorem should be applied to $\Omega = \Omega_{\text{feas}} \cap P$, since the omission of an impermissible alternative from S is not a defect of the procedure even if that alternative would have the highest value. A separate question, which the theorem does not answer, is whether the construction of P is itself legitimate. Optimization is thus conditional not only on what was considered but on which alternatives were admitted as candidates in the first place.

The second point concerns the dynamics of an institution that optimizes a measured objective. Let the state of the institution be $x_t = (k_t, v_t)$, where k_t represents capital and v_t represents the capacity for verification, which includes oversight, audit, and the competence to detect error. Let the dynamics be $x_{t+1} = F(x_t, a_t)$ and let the institution choose a policy π to maximize

$$J_m(\pi) = \mathbb{E}_\pi\Big[\sum_{t=0}^{T} \gamma^t \, r(k_t, a_t)\Big], \tag{11}$$

where the reward r depends on measured outputs but not on v_t. Verification capacity is distinct from productive competence. It does not change what is produced; it determines whether the

institution's errors are detected and corrected. Suppose that one unit of baseline error arises in each period, that verification catches a fraction v_t of it, and that each uncorrected unit costs $\kappa > 0$ of realized welfare which the measured reward does not record. Realized welfare in period t is then $r_t - \kappa(1 - v_t)$, so the gap between measured reward and realized welfare is $\kappa(1 - v_t)$ and widens as verification degrades.

Let the verification capacity evolve as

$$v_{t+1} = \min\{1, \max\{0,\, v_t - d(a_t) + s(a_t)\}\}, \tag{12}$$

where $d(a) \geq 0$ is the degradation of verification induced by action a and $s(a) \geq 0$ is investment in verification. The central question is whether an optimizer can preserve the epistemic conditions under which its own objective remains meaningful.

Proposition 12 (Endogenous verification decay)**.** *Suppose that neither the reward r nor the transition of the variables on which r depends involves v_t, and that investment is costly, in the sense that r is strictly decreasing in $s(a)$ when the other components of the action are held fixed and $s(a)$ can be set to zero without changing them. Suppose also that investment neither relaxes any constraint on the other components of the action nor changes the set of actions feasible in later periods. Then every policy that maximizes J_m over periods $t = 0, \ldots, T$ sets $s(a_t) = 0$ in every period; in particular, an investment made in the final period, which affects only v_{T+1} outside the horizon, is strictly dominated as well. If in addition the maximizing actions satisfy $d(a_t) \geq d_{\min} > 0$, then $v_t \leq \max\{0, v_0 - t\, d_{\min}\}$, so that $v_t = 0$ for every $t \geq t^* = \lceil v_0/d_{\min} \rceil$ provided $T \geq t^*$, and from then on the gap between measured reward and realized welfare equals κ in each period. If $T < t^*$, the proposition asserts only that verification falls by at least $d_{\min}$ in each period until it reaches zero.*

Proof. Because v_t enters neither r nor the dynamics of the variables that determine r, its value is irrelevant to J_m. A policy with $s(a_t) > 0$ in some period is therefore strictly dominated by the policy that is identical except that $s(a_t) = 0$ there, since this raises r and changes nothing else that J_m depends on. With $s = 0$ and $d \geq d_{\min}$, we have $v_{t+1} \leq \max\{0, v_t - d_{\min}\}$, which gives the stated bound by induction, and once $v_t = 0$ it remains zero because $v_{t+1} = \max\{0, -d(a_t)\} = 0$. The gap $\kappa(1 - v_t)$ then equals κ. □

The proposition identifies the structural condition under which decay is forced, which is that verification is not priced into the objective. If verification capacity affects future measured reward through the state, then it is valued instrumentally and decay is not forced. The remedy is therefore either to price verification into the objective or to protect it by constraint, and in both cases the commitment comes from outside V. A further extension would let verification determine the accuracy of the reward estimates available to later optimization, so that the optimizer progressively loses the ability to recognize its own mistakes although the production process is unchanged. Under these assumptions the conclusion is close to built into the model, and the proposition is best read as a demonstration of what follows when verification is excluded from the objective and from every channel through which it could become instrumentally

valuable. The next result relaxes this by letting verification have genuine instrumental value that the optimizer cannot correctly perceive.

Proposition 13 (Perceived-value trap)**.** *Suppose each unit of verification investment raises v by one unit at cost $c > 0$ and yields a true welfare benefit $\kappa > c$, net of discounting, so that investment is worthwhile at every level of v. Suppose the optimizer perceives that benefit as $\kappa\,\nu(v)$, where ν is continuous and strictly increasing on $[0,1]$ with $\nu(0) = 0$ and $\nu(1) = 1$, because only errors that verification has already made visible register as losses. Let $v^* = \nu^{-1}(c/\kappa) \in (0,1)$, and suppose the optimizer invests exactly when the perceived benefit is at least the cost. Then it invests nothing whenever $v < v^*$. If in addition degradation satisfies $d(a_t) \geq d_{\min} > 0$ whenever $v_t < v^*$, then verification reaches zero in finitely many periods and never recovers, although investment is worthwhile at every level.*

Proof. Since $c/\kappa \in (0,1)$ and ν is a strictly increasing bijection of $[0,1]$, the level v^* exists and lies in $(0,1)$. For $v < v^*$ the perceived benefit satisfies $\kappa\nu(v) < \kappa\nu(v^*) = c$, so the optimizer does not invest. With no investment and $d \geq d_{\min}$, $v_{t+1} \leq \max\{0, v_t - d_{\min}\} < v^*$, so the condition persists, and v_t reaches zero within $\lceil v_t/d_{\min} \rceil$ further periods and remains there. □

The proposition shows how an optimizer can rationally underinvest in verification that is genuinely valuable, because the value is visible only through the errors that verification has already exposed. A shock that pushes verification below v^* is then self-reinforcing. A two-period instance shows the divergence between measured and realized value.

Example 2 (Divergence of measured and realized value)**.** Let $T = 1$ and let there be two policies. Under *maintain*, the reward is 1 in each period and verification stays at $v_t = 1$, so there is no gap. Under *divert*, the institution moves effort from verification to production, obtaining a reward of $1 + \Delta$ in period 0 and driving v_1 to zero, so that in period 1 the reported reward is 1 while a unit of error goes uncorrected. The measured objectives are $J_m(\text{maintain}) = 1 + \gamma$ and $J_m(\text{divert}) = 1 + \Delta + \gamma$, so the measured objective prefers diverting effort from verification for every $\Delta > 0$. The realized welfare is $1 + \gamma$ under maintain and $1 + \Delta + \gamma(1 - \kappa)$ under divert, so when $0 < \Delta < \gamma\kappa$ the realized welfare is lower under the policy that the measured objective selects.

The example is stylized and its point is structural. This is not merely Goodhart's observation that a measure used as a target ceases to be a good measure. In the example the optimization changes the reliability of the information on which later optimization depends, and an institution that has degraded its own capacity to verify may progressively lose the ability to detect that its actions are no longer beneficial. The divergence between measured and realizable returns is therefore not constant and does not average out. It is generated by the optimization, and it is invisible to the measure that guides it. The remedy is to include verification capacity in the state and to treat its preservation as a constraint, and this again requires a commitment that the objective V does not supply.

11 Scalar Metrics as Decision Rules: A Case from AI Governance

The formal structure examined so far is not specific to philanthropy. It arises whenever a scalar quantity serves both as a predictor of something valued and as the basis of a decision to be coordinated among several parties, and a recent proposal in the governance of frontier AI development provides a clear instance. The two problems are not identical optimization problems. The earlier sections concern maximizing expected welfare over interventions, whereas this section concerns the use of predictive performance as a proxy for hazardous capability and as a trigger for a decision. What they share is a logical form, in which a scalar quantity is computed under a particular framing and then taken to license a conclusion about a larger decision problem, and the failure mechanisms differ accordingly: omitted alternatives and aggregation scale in the first case, the evaluation distribution and unmeasured routes to capability in the second. This section is best read as an application of the certification principle to a different class of decision problem, a classification rule based on a proxy measurement, and not as a claim that every scalar objective suffers the same pathology. Pachocki (2026) reports the perplexity, or equivalently the cross-entropy loss, of successive models on held-out data from an internal codebase, using a fixed snapshot. The post associates declining log-perplexity with the appearance of capabilities, from fluent conversation at roughly 0.99 to productive chain-of-thought reasoning at roughly 0.91, to transformative capabilities in cybersecurity and mathematics below 0.8 and results on open mathematical problems near 0.76. It anticipates further transformative capabilities in a window of roughly 0.68 to 0.70, and it proposes that developers commit to common limits based on perplexity and other measurements, not to be crossed until the ability to understand and monitor model behavior reaches a stated milestone. The post lists its own limitations: comparison across developers is difficult, perplexity does not map directly to tasks of interest, it ignores progress outside pretraining, and the floor of achievable loss is unknown. It also states that chain-of-thought reasoning raises capability without lowering the perplexity a model achieves. The analysis below does not dispute the measurements. It asks what each of the post's three claims requires.

The three claims have different evidentiary burdens. The first is empirical: reductions in base-model loss have historically accompanied the appearance of capabilities, which historical measurement can support. The second is extrapolative: further reductions will accompany further transformative capabilities, which requires a model of how capabilities emerge. The third is institutional: the loss can serve as a common measure for deciding when development should pause, which requires a justification for using the measure to coordinate decisions. The transitions among these claims are where the structure of the earlier sections reappears.

Let θ range over a class Θ of models, let q_θ be the model's predictive distribution, and for an evaluation distribution P let

$$L_P(\theta) = -\mathbb{E}_{x\sim P} \ln q_\theta(x), \tag{13}$$

so that perplexity is $\exp L_P(\theta)$. This is a measure of predictive performance relative to a distribution and not an intrinsic measure of intelligence, and the reported series is therefore

more precisely a series of values of L_P for the particular distribution P of that codebase snapshot. The ranking of models can depend on that choice.

Proposition 14 (Dependence of the ranking on the evaluation distribution)**.** *Let $P_w = wP_1 + (1-w)P_2$ for $w \in [0,1]$, and let L_i denote the loss under P_i. If model A has lower loss than model B under P_1 and higher loss under P_2, then there is $w^* \in (0,1)$ such that $L_{P_w}(A) < L_{P_w}(B)$ if and only if $w > w^*$.*

Proof. Write $d_i = L_i(A) - L_i(B)$, so that $d_1 < 0 < d_2$. Then $L_{P_w}(A) - L_{P_w}(B) = wd_1 + (1-w)d_2$ is strictly decreasing in w, positive at $w = 0$ and negative at $w = 1$, and it vanishes at $w^* = d_2/(d_2 - d_1) \in (0,1)$. □

There is no contradiction here, since one model may predict code better while another predicts prose or scientific text better. The mixture weight plays the role that the interpersonal scale played in Proposition 2, and the choice of evaluation distribution plays the role of the candidate set in Section 5. A shared threshold therefore requires a shared definition of what is measured, and the choice of distribution becomes part of the governance standard and not a technical preliminary to it.

A second difficulty is that the loss does not determine capability. Let $R(\theta)$ be a measure of a particular hazardous capability.

Proposition 15 (Insufficiency of scalar loss)**.** *If there exist $\theta_1, \theta_2 \in \Theta$ with $L_P(\theta_1) = L_P(\theta_2)$ and $R(\theta_1) \neq R(\theta_2)$, then no function f satisfies $R(\theta) = f(L_P(\theta))$ for all $\theta \in \Theta$. Moreover, if capability depends on an inference-time procedure κ such as reasoning at inference, tool access, or post-training, so that $R = R(\theta, \kappa)$ while L_P depends on θ alone, then any θ for which $R(\theta, \cdot)$ is not constant yields such a pair.*

Proof. A function cannot assign two values to one argument. For the second statement, take κ_1, κ_2 with $R(\theta, \kappa_1) \neq R(\theta, \kappa_2)$; the two systems share the loss $L_P(\theta)$. □

The proposition shows that loss is not a sufficient statistic for capability under the stated condition. It does not show that loss is useless for forecasting. A simple illustration of how average loss can conceal differences uses two equally weighted classes of problems. A model with class losses $(0.2, 1.4)$ and a model with class losses $(0.8, 0.8)$ both have average loss 0.8, while their behavior on the first class, which might contain the safety-critical tasks, differs greatly. The quantity of practical interest is the variation in capability among models of nearly equal loss,

$$\Delta_R(\ell, \varepsilon) = \sup_{\substack{\theta_1, \theta_2 \in \Theta \\ |L_P(\theta_i) - \ell| \leq \varepsilon}} \big|R(\theta_1) - R(\theta_2)\big|, \tag{14}$$

which can in principle be estimated by comparing architectures, post-training procedures, inference budgets, and tool access while holding base-model loss approximately fixed. If Δ_R is large near a proposed threshold, the threshold cannot reliably separate systems of substantially different capability.

The threshold claim should be separated into three propositions of increasing strength. That a lower loss improves average predictive performance on a fixed distribution follows from the definition. That a lower loss is accompanied by improvement in some downstream capabilities is an empirical regularity requiring evidence. That loss below a particular value L^* produces a particular hazardous capability is a much stronger claim, requiring a validated relationship between the threshold and the capability, and a forecast of a window in which transformative capabilities are expected is not a demonstration of it. There is also a selection effect of the kind identified in Section 5. The milestones annotated on a historical series, such as fluent conversation, useful coding, and mathematical discovery, are heterogeneous achievements noticed after they appeared, and the loss at which each was first noticed can be described afterward as a threshold. Whether such a threshold could have been predicted beforehand is a question for prospective testing with capability definitions fixed in advance. The association itself is observational in the sense of Section 7, since successive models differ in data, scale, and post-training at once, so the effect of lowering loss while holding the other factors fixed is not identified by the series alone.

The loss also has an elementary decomposition,

$$L_P(q) = H(P) + D_{\mathrm{KL}}(P \,\|\, q), \tag{15}$$

where $H(P)$ is the entropy of the evaluation distribution and the divergence measures the shortfall of the model from it (Cover and Thomas, 2006). Reducing the loss reduces the divergence. It does not say which internal representations changed, whether a given reasoning task improved, or whether the model acquired a general procedure and not better statistical coverage. When the loss is near the unknown floor $H(P)$, a numerically small decrease may be a large proportional reduction of the remaining divergence, but whether that produces a large increase in useful capability is a further empirical question, and a shrinking numerical gap does not by itself imply accelerating capability.

Now consider the institutional claim as a decision rule. Let a deployed system be a pair (θ, κ) of a base model and an inference-time configuration drawn from a set K, with capability $R(\theta, \kappa)$. Let $\mathrm{Pause}(\theta) = \mathbf{1}\{L_P(\theta) < \tau\}$ for an agreed threshold τ, so that the rule depends on the base model alone, and let r^* be a level of hazardous capability that the rule is meant to catch. Define

$$\omega_\tau = \sup\{R(\theta, \kappa) : \theta \in \Theta,\ \kappa \in K,\ L_P(\theta) \geq \tau\}, \tag{16}$$

with the convention that the supremum over an empty set is $-\infty$, so that the rule trivially detects every hazardous model when no model lies at or above τ.

Proposition 16 (Complete detection by a threshold rule)**.** *The rule flags every deployed system with $R(\theta, \kappa) > r^*$ if and only if $\omega_\tau \leq r^*$.*

Proof. A hazardous system escapes the rule exactly when $L_P(\theta) \geq \tau$ and $R(\theta, \kappa) > r^*$, and such a system exists if and only if $\omega_\tau > r^*$. □

The condition has the form of the omission term in Theorem 1. The models above the threshold, and the routes to capability that the loss does not register, play the part of the

unevaluated alternatives, and complete detection requires a bound on the supremum of capability over that region. A high conditional probability of hazard below the threshold would support additional scrutiny, but it is the behavior above the threshold that determines whether the rule is a safeguard, and the post's own observation that reasoning at inference raises capability without lowering loss is a mechanism by which ω_τ can be large. Complete detection of models exceeding a stated capability level is also weaker than the safety of the development or deployment process, which depends on what is done with the flags.

Completeness is likewise not the same as usefulness. A threshold that flags every model detects every hazardous model while discriminating nothing, so the proposition characterizes sensitivity and not the quality of the rule. A second quantity is needed, namely the cost of pausing models that fall below the hazard level, and an expected-cost comparison $C(\tau) = c_{\text{miss}}\, P(L_P \geq \tau,\, R > r^*) + c_{\text{pause}}\, P(L_P < \tau,\, R \leq r^*)$ expresses the trade-off. Because the flagged set grows with τ, the quantity ω_τ is nonincreasing in τ and the second term is nondecreasing, so among thresholds that achieve complete detection the smallest is the least costly in unnecessary pauses. Whether complete detection is worth that cost depends on the two cost parameters, which the loss does not determine. This comparison holds the distribution of systems fixed; once developers respond strategically to the threshold, that distribution changes and the comparison is no longer automatic. Here R in the cost expression denotes the capability of the deployed system. A governance threshold also creates an incentive. A developer who maximizes capability subject to $L_P(\theta) \geq \tau$ is constrained only along the dimension that the metric registers, so effort moves toward those that it does not, in the manner described in Section 8 for selection on a reported quantity.

These observations should be balanced against the strengths of the proposal. A loss measured on held-out data may be harder to manipulate through targeted elicitation than a changing collection of downstream benchmarks, and the proposal pairs the loss with other measurements and with a monitoring milestone, so it is not a rule based on perplexity alone. That is consistent with the thesis of this essay and not an objection to it: the scalar measure is informative, a procedure that is safe to act on requires further components, and whether the measure is sufficient to fix the threshold depends on empirical bounds on Δ_R near τ that the historical series does not supply. Two senses of perplexity now appear in the essay and should be kept apart. In Proposition 4 it summarizes a distribution over where the optimum lies, and here it is the exponential of a language model's cross-entropy; the mathematical form is shared, while the distributions and the questions are different.

12 What the Repairs Cost

The difficulties above have standard remedies, and each deserves to be assessed for what it changes. Bounded utility, $U \leq \bar{U}$, blocks Proposition 1, since then $pB \leq p\bar{U}$ and a certain harm H is outweighed only if $p\bar{U} > H$. The bound is now a parameter on which the recommendation depends, and the manner of implementation matters. A bounded transformation of total welfare implies that the value of a further life declines as the total grows, which conflicts with the

impartial additive view on which the largest benefits arose, whereas a cap on each individual's welfare leaves the total unbounded in the number of persons and so does not block Proposition 1 when the population is itself unbounded. The trilemma of Beckstead and Thomas (2024) shows that avoiding recklessness by this route incurs timidity or another cost, so the repair does not remove the commitment but moves it.

Robust criteria, such as maximin over a set of priors (Gilboa and Schmeidler, 1989), block domination whenever the set includes a probability of zero, and they do so at the price of widespread conservatism. If long-run consequences are uncertain for every action including inaction, as the cluelessness literature maintains (Greaves, 2016; Mogensen, 2021), then the maximality rule leaves many options incomparable and the criterion alone does not select among them, which is the practical incompleteness noted by Herlitz (2019). Side constraints that restrict the permissible action set, in the manner of Nozick (1974), restore the distinction between maximizing and permission, and they do so by introducing a second source of normative content that is not derived from V. Verification requirements address the dynamics of Section 10 and the claims of Section 8, and they consume resources whose allocation is itself subject to the recursion of Section 9.

The question that matters is whether the repairs can be combined without destroying the maximizing objective. Let C denote a family of constraints specifying the admissible actions, the admissible probability sets, the bound on utility, and the verification requirements. The decision procedure is then a map $(V, S, C) \mapsto a^*$, where $a^* = \arg\max_{a \in S \cap P_C} V(a)$ is evaluated under the representation of uncertainty that C fixes. Two procedures with identical V and different C recommend different actions, and nothing in V selects between them. This is the content of the conclusion that expected good is insufficient to specify a legitimate decision procedure. It follows from the structure of the problem, and it is entirely compatible with the view that expected good is a central consideration within a procedure that has other components.

13 Conclusion

The results of this essay are the following. An additive expected-value rule with unbounded benefits cannot exclude the conclusion that a large enough hypothetical benefit outweighs a certain harm (Proposition 1). A sum of individual welfare representations is not determined without interpersonal comparability assumptions (Proposition 2). The maximizer over a restricted candidate set is certified as globally optimal only under a bound on the excluded alternatives (Proposition 3 and Theorem 1), and enlarging the set can raise the selection bias of the winning estimate (Proposition 7). Interventions have context-independent marginal values only under separability, and rankings of single interventions can mislead when it fails (Proposition 9 and Example 1). Observed success under an existing allocation does not identify the causal ranking of interventions (Proposition 10). Selection on reported cost selects for favorable reporting error (Proposition 11), and an institution that optimizes a measured objective can degrade the verification on which the measure rests (Proposition 12 and Example 2). The

same structure appears when a scalar predictive loss is proposed as a threshold for coordinating decisions: the ranking of models depends on the evaluation distribution (Proposition 14), average loss does not determine capability (Proposition 15), and a rule gated on the loss detects every hazardous model only under a bound on capability among models above the threshold (Proposition 16).

What has not been shown deserves the same clarity. None of the results establishes that effective altruism is internally inconsistent, that useful comparisons are impossible, or that the movement's practitioners endorse harmful conduct. Each result is conditional on identified assumptions, and in each case the assumption can be satisfied or the repair adopted. The finding is that the quantitative apparatus, taken by itself, leaves undetermined several things that its recommendations depend on: the candidate set, the model of interaction, the identification of causal effects, the aggregation scale, the treatment of ambiguity, and the permission to act. These are the places where the arithmetic gives way to judgment, and a claim that an intervention is the most effective way of doing good is only as strong as the defence of those judgments. Nor does anything in Section 11 show that perplexity lacks predictive value; the claim is that predictive usefulness and sufficiency for a governance decision are different properties.

A natural continuation is a formal treatment of counterfactual selection under partial identification, in which the evaluator is uncertain not only about the values of known alternatives but also about the structure, existence, and potential value of alternatives outside the candidate set, together with a characterization of when a ranking remains valid under expansion of the candidate set for particular classes of search processes. The distinction between ranking uncertainty and omission uncertainty suggests that reports of cost-effectiveness should state the second as explicitly as the first, for example through the pair formed by the perplexity of the distribution over the evaluated candidates and the probability that the optimum lies outside them.

References

Arrhenius, G. (2000). An Impossibility Theorem for Welfarist Axiologies. *Economics and Philosophy*, 16(2), 247–266.

Beckstead, N., and Thomas, T. (2024). A Paradox for Tiny Probabilities and Enormous Values. *Noûs*, 58(2), 431–455.

Bostrom, N. (2003). Astronomical Waste: The Opportunity Cost of Delayed Technological Development. *Utilitas*, 15(3), 308–314.

Côté, N., and Steuwer, B. (2023). Better Vaguely Right than Precisely Wrong in Effective Altruism: The Problem of Marginalism. *Economics and Philosophy*, 39(1), 152–169.

Cover, T. M., and Thomas, J. A. (2006). *Elements of Information Theory*, 2nd ed. Hoboken: Wiley.

d'Aspremont, C., and Gevers, L. (1977). Equity and the Informational Basis of Collective Choice. *Review of Economic Studies*, 44(2), 199–209.

Efron, B. (1965). Increasing Properties of Pólya Frequency Functions. *Annals of Mathematical Statistics*, 36(1), 272–279.

Ellsberg, D. (1961). Risk, Ambiguity, and the Savage Axioms. *Quarterly Journal of Economics*, 75(4), 643–669.

Gilboa, I., and Schmeidler, D. (1989). Maxmin Expected Utility with Non-Unique Prior. *Journal of Mathematical Economics*, 18(2), 141–153.

Greaves, H. (2016). Cluelessness. *Proceedings of the Aristotelian Society*, 116(3), 311–339.

Greaves, H., Thomas, T., Mogensen, A., and MacAskill, W. (2024). On the Desire to Make a Difference. *Philosophical Studies*, 181(6), 1599–1626.

Harsanyi, J. C. (1955). Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility. *Journal of Political Economy*, 63(4), 309–321.

Herlitz, A. (2019). Cluelessness and Rational Choice: The Case of Effective Altruism. *Ethical Perspectives*, 26(3), 401–426.

Howard, R. A. (1966). Information Value Theory. *IEEE Transactions on Systems Science and Cybernetics*, 2(1), 22–26.

Jelinek, F., Mercer, R. L., Bahl, L. R., and Baker, J. K. (1977). Perplexity: A Measure of the Difficulty of Speech Recognition Tasks. *Journal of the Acoustical Society of America*, 62(S1), S63.

Lenman, J. (2000). Consequentialism and Cluelessness. *Philosophy and Public Affairs*, 29(4), 342–370.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *American Economic Review*, 80(2), 319–323.

Mogensen, A. L. (2021). Maximal Cluelessness. *Philosophical Quarterly*, 71(1), 141–162.

Nozick, R. (1974). *Anarchy, State, and Utopia*. New York: Basic Books.

Pachocki, J. (2026). Perplexity and Intelligence. OpenAI, post of 22 November 2026. `https://openai.com/index/perplexity-and-intelligence/`

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference*, 2nd ed. Cambridge: Cambridge University Press.

Savage, L. J. (1954). *The Foundations of Statistics*. New York: Wiley.

Sen, A. (1970). *Collective Choice and Social Welfare*. San Francisco: Holden-Day.

Smith, J. E., and Winkler, R. L. (2006). The Optimizer's Curse: Skepticism and Postdecision Surprise in Decision Analysis. *Management Science*, 52(3), 311–322.

United States Department of Justice, Office of Public Affairs. (2024). Samuel Bankman-Fried Sentenced to 25 Years for His Orchestration of Multiple Fraudulent Schemes. Press release, 28 March 2024.

Wilkinson, H. (2022). In Defense of Fanaticism. *Ethics*, 132(2), 445–477.