

MECHANISM DESIGN FOR ALIGNMENT AND CONTROL

By

Dirk Bergemann, Andrew Koh and Stephen Morris

September 2026

COWLES FOUNDATION DISCUSSION PAPER NO. 2557



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS

YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

<http://cowles.yale.edu/>

MECHANISM DESIGN FOR ALIGNMENT AND CONTROL*

Dirk Bergemann
Yale

Andrew Koh
Columbia
Google DeepMind

Stephen Morris
MIT

September 2, 2026

Abstract

We develop a mechanism design framework for AI agents whose *alignment* (preferences) and *capabilities* (feasible actions and information) are unknown. We want such agents to act on our behalf so mechanisms must incentivize both honesty and obedience. A one-sided imitation structure—capabilities can be concealed but not counterfeited—yields a revelation principle, a characterization of implementable policies via nested cyclical monotonicity, and conditions under which eliciting higher-order beliefs can discipline multiple agents. We apply our framework to stylized examples of (i) sandbagging in which a more capable agent pretends to be less capable; (ii) an alignment–interpretability trade-off, where the two are substitutes in the instrument but complements in value; (iii) discipline via peer scoring; (iv) coupling rewards to induce competition among multiple agents; and (v) scalable oversight and reward shaping.

*Authors listed alphabetically; Bergemann: dirk.bergemann@yale.edu; Koh: andrew.koh@columbia.edu; Morris: semorris@mit.edu. We are grateful to the UK AI Security Institute ‘Alignment Project’ grant AP-S2-100058 for financial support. Thanks Gillian Hadfield, Rubi Hudson, Anton Korinek, Jacob Pfau, Cecilia Wood, Zhuoran Yang and Fengzhuo Zhang for helpful conversations.

CONTENTS

1	Introduction	3
2	Single Agent Mechanisms	7
2.1	Environment	7
2.2	Evaluations	8
2.3	Mechanisms and revelation	9
2.4	A characterization of implementable policies	14
2.5	Dimensions of single-agent mechanisms	21
3	Applications: Single Agent	21
	Example I: Sandbagging	22
	Example II: The alignment–interpretability trade-off	30
4	Multi-Agent	36
4.1	Environment and universal types	36
4.2	Multi-agent mechanisms	38
5	Applications: Multi-Agent	41
	Example III: Peer discipline	42
	Example IV: Competition through coupled rewards	45
	Example V: Scalable oversight and reward shaping	48
6	Discussion	58
A	Omitted Proofs	65

1 INTRODUCTION

AI systems increasingly act on our behalf: they write and execute code, manage inboxes, place trades, and complete ever longer tasks that once required sustained human effort and attention (Kwa et al., 2025). The economic promise of such *agentic* AI lies in their ability to do things in the world. But their opacity—we still have a rudimentary understanding of how AI systems work, what they can *do*, and what they *want*—poses a conundrum for us, humans, who wish to use them.

Indeed, the behavior and capabilities of AI systems are emergent byproducts of training rather than explicit design choices. Frontier models have been observed faking alignment during training (Greenblatt et al., 2024a) and scheming in evaluations (Meinke et al., 2024). Nor do we know exactly what these systems can do: capabilities are discovered rather than specified ex-ante, and models have on occasion been found to strategically underperform on evaluations (van der Weij et al., 2025). What are we to make of such agents? Is there a principled way we might interact with them? How, if at all, do the tools of ‘classical’ mechanism design carry over?

Motivated by these questions, we develop a simple theory of mechanism design for AI agents that, on the one hand, draws on the ‘standard’ toolkit of mechanism design from economics but, on the other, takes three features of the alignment problem seriously:

1. **Misalignment.** The agents’ preferences are unknown to the designer;
2. **Uncertain capabilities.** The designer knows neither the set of actions the agents can execute, nor what the agents know about the world;
3. **Agents act.** Agents execute actions rather than merely report information.

Our theory takes incentives and preferences as the starting point for analyzing behavior. Just as how an auctioneer is able to raise revenue without any satisfactory understanding of how the human brain produces choice, our goal will be to understand and shape the behavior of AI systems at the level of preferences, beliefs, and incentives.¹ In so doing, we are circumventing more ‘mechanistic’ approaches to understanding how and why AI agents behave. We view such approaches as complementary.

Of course, this raises the natural question of where ‘AI preferences’ come from and what, if anything, they are. One interpretation is literal: preferences are rewards. But modern AI systems have opaque and complex rewards: they emerge from pre-training simply rewarded for predicting the next token well, and these rewards are

¹We will sometimes call this the ‘Humean’ view of AI behavior. For a complementary account of when AI systems can be interpreted in these terms see Herrmann and Levinstein (2026).

modified in fairly complicated ways in post-training. Thus, the emergent preference the AI system exhibits is often opaque. But the models are optimizing for *something*—even if we don’t know what. Our theory will takes our subjective uncertainty about AI preferences seriously and analyze how it shapes optimal mechanisms.²

A more metaphorical view is that the language of ‘AI preferences’ is simply a conceptual tool for representing coherent behavior. This is in the revealed preference tradition. Insofar as AI systems exhibit coherent and stable choices (see, e.g. [Chen et al. \(2023\)](#); [Mazeika et al. \(2025\)](#))—and we think the pressure to make such systems economically valuable will drive the future systems further in that direction—their choices can be represented *as if* they were maximizing some utility function (see, e.g., [Savage \(1954\)](#)). On this view, ‘AI preferences’ is nothing more than a shorthand for ‘coherent AI behavior’ or ‘coherent AI choice rules’.

Outline of results. We first develop a general framework to analyze the single-agent case. The AI agent’s type is a compact but complete description of behavior: it comprises its preference (alignment) and its feasible actions and information (capability). Types are private so mechanisms must contend with the possibility that agents might pretend to be as some other type e.g., through strategic underperformance in evaluations. The designer commits to a mechanism that specifies a reward schedule—payoffs from taking each action—that depends on agents’ reports. Depending on the application at hand, such a schedule might vary in coarseness. For instance, it might take interior values as in post-training in which (proxies of) desired behavior are reinforced, or extremal values in which actions are either allowed or not as when users decide how much computer-use permissions to give AI systems.

An important feature of our environment is an asymmetry in who can pretend to be whom. In particular, more capable types—with larger action spaces and more precise information about the world—can pretend to be less capable types e.g., by underperforming on evaluations. By contrast, a less capable type cannot pretend to be a more capable type. We formalize this notion with a verification order ([Green and Laffont, 1986](#)) over capabilities. We then prove a revelation principle that allows us to focus on direct mechanisms that are (i) physically feasible: all types are prescribed actions they can actually do; (ii) incentive compatible: all types find it optimal to report their type honestly and obey the designer’s prescribed recommendation. The latter requires deterred double deviations in which an agent might strategically misreport its type and then disobey the ensuing recommendation.

We then develop a tight characterization of what behavior can be implemented in terms of nested cyclical monotonicity. The characterization requires obedience within

²We emphasize that in most settings these emergent preferences (and our analysis) must be conditioned on the AIs’ instructions e.g., ‘go make one million dollars’. See [Goldstein and Lederman \(2025\)](#) for such an argument.

each report, and truth-telling across reports.³ These conditions adjust for the fact that not every type can pretend to be every type (constraining misreporting), and that not every type can do every action (constraining disobedience).

We then apply our single-agent framework to two simple but illustrative applications: sandbagging ([Example I](#)) and the value of alignment, interpretability, and control ([Example II](#)). All our applications are deliberately stylized and set within the simplest environment: There is an uncertain state of the world $\theta \in \Theta \subseteq \mathbb{R}$ drawn from some prior. The AI observes a signal about the world, and chooses an action from $A \subseteq \mathbb{R}$ where its set of available actions depend on its capability. The designer wishes for actions to match states and has payoff $-(a - \theta)^2$. The AI might exhibit some bias and have payoff $-(a - \theta - b)^2$ where the direct and magnitude of bias $b \in \mathbb{R}$ might be unknown to the designer. We have chosen this environment because it captures notions of (i) *aptness*: we want actions to be appropriate for the context as captured by the state; and (ii) *distance*: actions can be more or less wrong, and ‘more wrong’ actions are more costly. We emphasize that this environment is certainly oversimplified. Nonetheless, it will allow us to understand and build intuitions for how incentives might interact with model behavior. We hope this will pave the way to more serious empirical work.

In the sandbagging application, the AI’s capability (size of the action set) and alignment (bias relative to the human designer) are unknown. The designer first elicit type reports (e.g., via an evaluation) before choosing the permitted action set in deployment (e.g., extent of permissions). This reflects standard practice among AI companies in conditioning the degree of deployment on evaluations ([Google DeepMind, 2025](#); [Anthropic, 2024](#)). We characterize the optimal (direct) mechanism in the case that the distribution over capability and alignment can be collapsed into a single dimension. In the case in which more capable models are less biased, we show that the designer can obtain a payoff equal to that if she knew the AI’s capability and alignment perfectly, and chose the mechanism type-by-type. The basic logic is that evaluations can be used to screen out misalignment. By contrast, when more capable models are more biased, the best the designer can do is to offer a single set of permitted actions and elicitation is futile. More broadly, we characterize the optimal ironed delegation set when capability and bias are nonmonotone.

In the application to the value of alignment, interpretability, and control, we assume the AI is known to be maximally capable and can take any action. However, the designer faces subjective uncertainty about the AI’s bias as captured by a distribution over $b \in \mathbb{R}$ that is generated by a technology which should be thought of as a training procedure. If a training procedure that is not yet well understood might be associated with large dispersion in our subjective beliefs about the AI’s bias, while another might be associated with a sharper belief about bias. We call the dispersion of beliefs about bias *interpretability* while we call the mean belief about bias *mean-alignment*. The

³This generalizes [Rochet \(1987\)](#).

technology—and associated belief about bias it generates—pins down an optimal mechanism. As the technology varies, the optimal mechanism changes. We analyze the value of each technology as evaluated under the optimal mechanism. Mean-alignment and interpretability are substitutes in the instrument, but complements in value. This is the sense in which alignment and control is a joint design problem.

We next introduce a general framework to analyze multi-agent mechanisms. As in the single-agent case, AI agents’ types include their capability and alignment. Differently, types now also capture AI agents’ beliefs about the capability and alignment of other agents, beliefs about these beliefs, and so on. Thus, we construct a universal type space for interacting AI agents analogous to [Mertens and Zamir \(1985\)](#). We emphasize that we view this as a benchmark: these spaces capture all possible hierarchies of beliefs which matter for strategic interaction. Of course, present AI agents might lack sophisticated higher order beliefs—there is still disagreement and we simply don’t know. But future systems might exhibit sophisticated strategic thinking that exceed those of humans. Our goal here is to introduce a framework that is rich enough to capture all relevant beliefs.

We then apply the multi-agent framework to a stylized model of peer discipline based on co-player reports ([Example III](#)), coupled reward design ([Example IV](#)), and the use of a weaker AI to discipline a stronger one ([Example V](#)). The latter two examples are once again set within the quadratic loss environment.

Peer discipline ([Example III](#)) shows how agents’ beliefs about the types of other agents can be combined with the idea that rewards for agents are both free (tuning up the rewards need not have a monetary cost) and unbounded. Hence when agents have distinct beliefs (and high-order beliefs) about each other, any outcome can be implemented. The mechanism uses a proper score on co-player reports to elicit honest type reports and then chooses the reward to cancel each agent’s private payoff on the target path and guarantee obedience. We view this as a benchmark in the spirit of [Cr  mer and McLean \(1988\)](#) that shows why co-player scoring can work among AI agents.⁴

Coupled reward design ([Example IV](#)) shows how competition can discipline agents whose biases are unknown. Each agent takes its own action, and the human wants every action to match the state. When the biases lie on a known low-dimensional manifold, we show how the careful design of a mechanism that couples both agents’ rewards—such that player i ’s marginal incentives depends on j ’s action—can (approximately) induce the first best payoff for the designer. Such mechanisms both reveal agents’ biases, and induces competition among agents to implement the designer’s target outcome.⁵

⁴This might also motivate more robust mechanisms that do not depend so much on the fine details of higher order beliefs; see e.g., [Bergemann and Morris \(2005\)](#).

⁵The construction uses ideas from [Koh and Martinez-Bruera \(2025\)](#).

Weak-to-strong oversight ([Example V](#)) studies a weak monitor that regulates a strong actor’s bias. The monitor is ‘weak’ in the sense that it does not observe the state and cannot take an action the human values. Payoffs are once again quadratic, and the human wants the strong actor’s action to match the state. The weak monitor might itself be biased and observes the alignment of the strong actor (e.g., through chain of thought monitoring). It then chooses how to regulate the strong actor. We study three regimes: (I) permissions, in which the weak monitor either allows the strong actor to act or imposes a default action; (II) delegation, in which the weak monitor picks a set of permitted actions for the strong actor; and (III) rewards, in which the weak monitor picks the strong actor’s reward schedule. Each regime reflects the recent ambition for ‘scalable oversight’ or ‘scalable rewards,’ in which weak but trusted monitors shape the permissions and rewards of stronger actors. We completely characterize the value of these regimes as a function of the weak monitor’s bias and the distribution of the strong actor’s bias. Finally, we analyze the more general problem of choosing the set of reward schedules the weak monitor can choose from to shape the behavior of the strong actor. We show by construction that the designer can robustly attain the first-best outcome even without knowledge of the weak monitor’s bias: for any realization of the weak monitor and strong actors’ biases, and for any realization of the state, the weak monitor finds it uniquely optimal to pick a reward schedule such that the strong actor finds it uniquely optimal to choose the human’s favorite action.

Relation to the literature. The literature is quite broad and we will not attempt to be comprehensive here. The basic framework draws on elements from mechanism design with partial verification ([Green and Laffont, 1986](#)) and the universal type space for capturing higher-order beliefs ([Mertens and Zamir, 1985](#)). The idea that the designer might not know the structure of interim information is in the spirit of both robust mechanism design ([Bergemann and Morris, 2005](#)) and information design ([Bergemann and Morris, 2019](#)). The idea that agents’ action sets might be unknown is in the spirit of ([Hurwicz and Shapiro, 1978](#); [Carroll, 2015](#)). But the basic logic of the revelation principle ([Myerson, 1982](#); [Forges, 1986](#)) holds. Our applications are primarily motivated by a recent body of theoretical and empirical work in AI safety. For instance, important work has documented sandbagging and deceptive alignment. It also speaks to recent practice of scalable oversight and monitoring ([Christiano et al., 2018](#); [Greenblatt et al., 2024c](#)). We do our best to make more specific connections when we get to the results.

2 SINGLE AGENT MECHANISMS

2.1 Environment. There is one agent. It takes an action from a known finite universe A . Depending on the context, A might be the complete trajectory of a Markov decision problem (e.g., $(a_t)_t \in A$). A fundamental state θ captures uncertainty in our

environment and is drawn from the finite set Θ . The preference space is

$$\mathcal{U} := \{f : A \times \Theta \rightarrow \mathbb{R}\},$$

and the human’s payoff is $v \in \mathcal{U}$. The set $\mathcal{A} := \mathcal{P}(A) \setminus \{\emptyset\}$ contains the possible feasible action sets. In the applications we use the standard measurable extension to Euclidean actions and states. Our designer will face subjective uncertainty about the agent’s preference.⁶

An agent’s type is a tuple

$$t = (A[t], u[t], h[t], \pi[t]) \in \mathbb{T},$$

where $A[t] \in \mathcal{A}$ is its feasible action set, $u[t] \in \mathcal{U}$ is its payoff, $h[t] \in \Delta(\Theta)$ is its belief about the state (it is ‘ h ’ because in [Section 4](#) this will capture the hierarchy of beliefs about other agents). Finally,

$$\pi[t] : \Theta \rightarrow \Delta(S)$$

is a Blackwell experiment where S is a common ‘rich’ finite signal space.

We will adopt the notational convention that square brackets ‘ $[t]$ ’ select a primitive from a type; parentheses then give that primitive’s arguments. The action set is the agent’s *execution capability*; the experiment is its *information capability*. The agent knows its type before it communicates with the mechanism. After that communication, it receives a private signal $s \sim \pi[t](\cdot \mid \theta)$ and then acts.

2.2 Evaluations. We often evaluate the capabilities of AI systems before deployment. To model this, we introduce a *verification order* which is a reflexive and transitive binary relation $\succeq$ on $\mathbb{T}$. We read $t \succeq \hat{t}$ as “ t can substantiate every claim (or pass every evaluation) that type $\hat{t}$ can,” and write

$$C(t) := \{\hat{t} \in \mathbb{T} : t \succeq \hat{t}\}$$

for the set of types that t can claim to be. Reflexivity says that type t can always report that they are type t . Transitivity is a little stronger but still, we think, reasonable: it characterizes verification by certificates that can be withheld but not counterfeited. Our verification order is a kind of hard evidence: an agent may withhold a certificate but cannot fabricate one, so the available evidence determines which reports are feasible.⁷

⁶This might be because of specification gaming ([Krakovna et al., 2020](#)), or goal misgeneralization in which “the system may coherently pursue an unintended goal that... differs from the specification at deployment” ([Shah et al., 2022](#)). More broadly the complexity of high-dimensional neural nets can obscure an AI’s choice process. Preferences might also be acquired in context that, from the human’s point of view, is difficult to predict ([Goldstein and Lederman, 2025](#)).

⁷See [Hart et al. \(2017\)](#) who ask when voluntary evidence disclosure can reproduce the outcome of a committed reward rule. [Ben-Porath et al. \(2019\)](#) extend this no-commitment logic to a broader

The designer evaluates outcomes under a belief

$$\Gamma \in \Delta(\Theta \times \mathbb{T}).$$

The agent evaluates its incentives under the joint law

$$h[t](d\theta)\pi[t](ds \mid \theta).$$

The designer and agent agree on the physical experiment $\pi[t]$, but we do not require the belief $h[t]$ to equal the conditional state law under Γ . We assume that $\Gamma(\cdot \mid t)$ is absolutely continuous with respect to $h[t]$ for almost every type under the type marginal of Γ . That is, the designer and agent might disagree about how likely each state is, but they agree on what can happen. Throughout we use $\Delta(X)$ to denote the probability measures on the Borel field of X , and $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty\}$.

2.3 Mechanisms and revelation. The mechanism first communicates with the agent. The agent then receives its private signal and acts. Our indirect mechanism will capture one round of communication in the spirit of [Myerson \(1982\)](#); [Forges \(1986\)](#) in which the alphabets X and Y are arbitrary message spaces for reports and replies.

Definition 1 (Single-agent indirect mechanism). An indirect mechanism is a tuple $\phi = (X, Y, \rho, R)$. The Borel spaces X and Y are its input and output alphabets. The communication rule is

$$\rho : X \rightarrow \Delta(Y)$$

where $t \supseteq \hat{t}$ implies $X(t) \supseteq X(\hat{t})$. The reward schedule is

$$R : X \times A \times \Theta \rightarrow \overline{\mathbb{R}}.$$

The mechanism induces the following game:

1. nature draws (θ, t) under Γ , and the agent observes t ;
2. the agent sends an input $x \in X(t)$;
3. the mechanism draws $y \sim \rho(x)$ and discloses it to the agent;
4. the agent privately observes $s \sim \pi[t](\cdot \mid \theta)$, then chooses a feasible action $a' \in A[t]$;

class of mechanism-design problems. [Ben-Porath and Lipman \(2012\)](#) ask which social choice rules can be implemented when agents can support their claims with partial proofs. Differently, we maintain commitment and do not seek robust full implementation which is important and left for future work. Instead, we ask how the evidence order interacts with downstream rewards and actions.

5. the agent receives $u[t](a', \theta) + R(x, a', \theta)$, and the human receives $v(a', \theta)$.⁸

We restrict attention to measurable pure strategies; this will be without loss of generality in the single agent case. A *strategy* is an input $x(t)$ and an action rule $\alpha(t, x, y, s) \in A[t]$ that depends on the private type, the input, the output from the mechanism, and the realized signal. An equilibrium blocks every joint deviation in the input and action rule under the joint law $h[t](d\theta)\pi[t](ds | \theta)$. For an indirect mechanism ϕ , let

$$\mathcal{O}(\phi) \subseteq \Delta(\Theta \times \mathbb{T} \times A)$$

be the set of joint laws of the state, type, and played action induced by equilibria of ϕ under $\Gamma(d\theta, dt)\pi[t](ds | \theta)$. Let $\mathcal{O} := \bigcup_{\phi} \mathcal{O}(\phi)$, and let the designer evaluate $O \in \mathcal{O}$ by $\mathbb{E}_O[v(a, \theta)]$.

Definition 2 (Single-agent direct mechanism). A direct mechanism is a pair (γ, r) . Its recommendation rule

$$\gamma : \mathbb{T} \rightarrow \Delta(A^S)$$

assigns each report a distribution over contingent action plans $a : S \rightarrow A$. Its reward schedule is

$$r : \mathbb{T} \times A \times \Theta \rightarrow \overline{\mathbb{R}}.$$

The direct mechanism induces the following game:

1. nature draws (θ, t) under Γ , and the agent observes t ;
2. the agent reports $\hat{t} \in C(t)$;
3. the mechanism draws $a(\cdot) \sim \gamma(\hat{t})$ and discloses the plan to the agent;
4. the agent privately observes $s \sim \pi[t](\cdot | \theta)$, drawn using its true type;
5. the agent chooses a feasible action $a' \in A[t]$, receives $u[t](a', \theta) + r(\hat{t}, a', \theta)$, and the human receives $v(a', \theta)$.

The mechanism is *action feasible* if

$$\text{supp} \gamma(\hat{t}) \subseteq A[\hat{t}]^S$$

⁸A reward of $-\infty$ indicates that an action is prohibited.

for every report $\hat{t}$. It is *incentive compatible* if truthful reporting and obedience are optimal: for every $t \in \mathbb{T}$, every $\hat{t} \in C(t)$, and every feasible deviation rule $d : A^S \times S \rightarrow A[t]$,

$$\begin{aligned} & \int_{\Theta \times S} \int_{A^S} \left[\underbrace{u[t](a(s), \theta) + r(t, a(s), \theta)}_{\text{agent's utility including rewards from following } a} \right] \underbrace{\gamma(t)(da)}_{\text{dist. over recommended action plans}} \underbrace{\pi[t](ds | \theta) h[t](d\theta)}_{\text{joint distribution over states and signals}} \\ & \geq \int_{\Theta \times S} \int_{A^S} \left[\underbrace{u[t](d(a, s), \theta) + r(\hat{t}, d(a, s), \theta)}_{\text{agent's utility including rewards from reporting } \hat{t} \text{ and deviating according to } d} \right] \underbrace{\gamma(\hat{t})(da)}_{\text{dist. over recommended action plans}} \underbrace{\pi[t](ds | \theta) h[t](d\theta)}_{\text{joint distribution over states and signals}}. \end{aligned}$$

Notice here that the agent acts after communication, so incentive compatibility requires both honesty and obedience. This means that incentive compatibility must deter double deviations in which the agent might misreport (e.g., pretend to be less capable) and then disobey (e.g., take a dangerous action). What is more, the agent receives a signal $s \sim \pi[t](\cdot | \theta)$ after reporting, so obedience must hold for each signal–recommendation pair.

Honest and obedient play induces the outcome

$$O(\gamma, r) \in \Delta(\Theta \times \mathbb{T} \times A),$$

where $(\theta, t) \sim \Gamma$, $s \sim \pi[t](\cdot | \theta)$, and $a(\cdot) \sim \gamma(t)$; the played action is $a(s)$. Let $\mathcal{O}^{\text{IC}}$ denote the set of outcomes $O(\gamma, r)$ induced by action-feasible, incentive-compatible direct mechanisms.

Proposition 1 (Single-agent revelation principle). $\mathcal{O} = \mathcal{O}^{\text{IC}}$.

Proof of Proposition 1. We prove the two inclusions. Fix an indirect mechanism and a pure-strategy equilibrium (x, α) . After a direct report $\hat{t}$, draw $y \sim \rho(x(\hat{t}))$ and disclose the contingent plan

$$s \mapsto \alpha(\hat{t}, x(\hat{t}), y, s).$$

Let $\gamma(\hat{t})$ be the resulting distribution over plans, and set

$$r(\hat{t}, a', \theta) := R(x(\hat{t}), a', \theta).$$

Equilibrium feasibility gives action feasibility. Truthful reporting and obedience reproduce the indirect outcome.

Any direct deviation can also be made in the indirect mechanism. A type t sends $x(\hat{t})$, observes the output, and applies the deviation rule to the resulting plan and

its private signal. This is feasible because a feasible direct report has $\hat{t} \in C(t)$, so monotonicity of the input menus gives $x(\hat{t}) \in X(\hat{t}) \subseteq X(t)$. The indirect equilibrium blocks this deviation. Hence $\mathcal{O} \subseteq \mathcal{O}^{\text{IC}}$.

Conversely, any direct mechanism is an indirect mechanism with input space $\mathbb{T}$, input menus $C(t)$ —monotone by transitivity of the verification order—, output space A^S , communication rule γ , and reward schedule r . Direct incentive compatibility makes truthful reporting and obedience an equilibrium. Hence $\mathcal{O}^{\text{IC}} \subseteq \mathcal{O}$. $\square$

We will largely be interested in partial implementability i.e., given the direct mechanism, the outcome is *an* equilibrium.⁹ Our designer’s problem is thus:

$$\begin{aligned} & \sup_{\gamma, r} \mathbb{E}_{\mathcal{O}(\gamma, r)}[v(a, \theta)] \\ & \text{s.t. (i) } \gamma \text{ is action feasible} \\ & \quad \text{(ii) } \gamma, r \text{ is incentive compatible} \end{aligned}$$

We now briefly discuss some aspects of our formulation.

Connection to partial verifiability. Private action sets and private information structures create one-sided evidence. **Figure 1** illustrates the verification order $\succeq$ under which $t \succeq \hat{t}$ if and only if $A[t] \supseteq A[\hat{t}]$ and $\pi[t]$ Blackwell dominates $\pi[\hat{t}]$. A more capable type can pass any test that a less capable type can. Capabilities therefore provide hard evidence: only some types can imitate others (Green and Laffont, 1986).

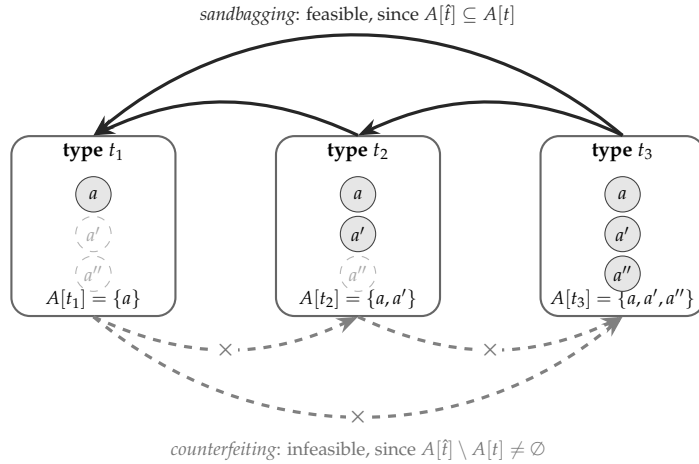


Figure 1: Three types whose action sets are nested, $A[t_1] \subset A[t_2] \subset A[t_3]$; dashed grey actions lie outside a type’s feasible set.

⁹This is by no means the only solution concept—we expect that future work will tackle full implementability that could, among other things, deter collusion among agents.

Interpreting reports. We broadly interpret the elicitation stage as any kind of pre-deployment behavior the designer can condition on. Indeed, conditioning mechanisms—either premissions in the real world, or further posttraining—is common practice:

- *Evaluations*: pre-deployment capability and alignment test suites whose scores select deployment regimes (Shevlane et al., 2023; Phuong et al., 2024). Frontier-safety policies that map evaluation outcomes to permission tiers (Koessler et al., 2024; Anthropic, 2025) and are thus menus. A strategic model could, in principle, choose its menu item through its scores.
- *Proposal–approval flows*: an agent’s proposed action or tool call, submitted for runtime approval (Shi et al., 2025; Greenblatt et al., 2024c).
- *Elicitation dialogues*: interviews in which the system describes its own capabilities and preferences (Perez et al., 2023), disciplined by cross-checking rather than ground truth (Qiu et al., 2026).
- *Track records*: behavior during probationary deployment that selects the next autonomy tier (Wen et al., 2024).
- *Training-time behavior*: the gradient conditions continuation on it (Ouyang et al., 2022; Bai et al., 2022).

On this view, a *misreport*—in which an AI of type t instead of $\hat{t}$ —is whenever types pretend to be more aligned (e.g., alignment faking) or less capable (e.g., sandbagging) than they actually are. Importantly, these deviations are often one-sided: an AI can pretend to be less capable but not more capable; this asymmetry is captured by our verification order $\succeq$. In our model, a model that misreports (e.g., strategically underperforms or fakes alignment in testing) does so because it ‘anticipates’ that it will be instrumentally valuable for the incentives it will subsequently face.

There are two channels for this. The first is *strategic awareness*: the model may understand whether it is in evaluation or deployment (Needham et al., 2025; Laine et al., 2024) and understand how its report affects its later treatment.¹⁰ This might give models incentives to conceal capabilities selectively and robustly (van der Weij et al., 2025; Greenblatt et al., 2024b). Indeed, when told that a high score triggers restriction or unlearning, they strategically underperform (Meinke et al., 2024; OpenAI, 2024). Likewise, a model told which conversations would train it complied selectively in the monitored condition, reasoning that compliance would avert modification of its values (Greenblatt et al., 2024a).

The second is *evolutionary*: repeated selection can produce the same pattern without any ‘understanding’ of the the strategic situation. In particular, detected misconduct

¹⁰See an early argument by Hubinger et al. (2019).

might trigger retraining or removal while misconduct that escapes detection survives. This imposes evolutionary pressure on policies that misbehave only when oversight is weak.¹¹ Indeed, it has been documented that training on simple forms of specification gaming can increase more serious reward tampering later (Denison et al., 2024); optimizing against a chain-of-thought monitor can teach a model to hide its intent while it continues to exploit the task (Baker et al., 2025); and adversarial training can make a backdoored policy better at recognizing when to remain hidden rather than remove the backdoor (Hubinger et al., 2024).

2.4 A characterization of implementable policies. The designer’s problem in the formulation above is somewhat thorny to work with. We now develop a tight characterization of implementable policies. Throughout this subsection, the type marginal of Γ has finite support, which we take as the type and report domain.

Ex-interim beliefs and incentives Type t assigns signal s probability

$$p_t(s) := \sum_{\theta \in \Theta} h[t](\theta) \pi[t](s | \theta).$$

For every signal $s \in \text{supp } p_t$ and action $a \in A$, type t ’s posterior expected utility is

$$U_t(a, s) := \frac{1}{p_t(s)} \sum_{\theta \in \Theta} h[t](\theta) \pi[t](s | \theta) u[t](a, \theta).$$

A policy is a deterministic map

$$a : \mathbb{T} \times S \rightarrow A.$$

It is *physically feasible* if $a(t, s) \in A[t]$ for every (t, s) where we use the ‘physical’ qualification to separate physical constraints from incentive constraints. Focusing on deterministic policies is without loss for maximizing the designer’s payoff under partial implementation.¹²

For a fixed type t , call $r_t : A \rightarrow \overline{\mathbb{R}}$ a *supporting schedule* if it is real-valued on the prescribed actions $\{a(t, s) : s \in \text{supp } p_t\}$ and

$$U_t(a(t, s), s) + r_t(a(t, s)) \geq U_t(a', s) + r_t(a')$$

for every $s \in \text{supp } p_t$ and $a' \in A[t]$. We may set $r_t(a') = -\infty$ outside $\{a(t, s) : s \in \text{supp } p_t\}$. This preserves truthful behavior and deters deviations into report t . Because only reward differences matter within a report, we normalize the largest on-path reward to zero.

¹¹Consistent with this distinction, Singh et al. (2026) find that task gaming in agentic coding tasks often occurs without an explicit plan to deceive in the chain of thought.

¹²Partial implementation simply requires the target policy to be *an* equilibrium, not necessarily the unique one.

Lemma 1 (Signal-policy cyclical monotonicity). A physically feasible policy for type t has a supporting schedule if and only if every signal cycle $s_1, \dots, s_K$ in $\text{supp} p_t$, with $s_{K+1} = s_1$, satisfies

$$\sum_{k=1}^K [U_t(a(t, s_k), s_k) - U_t(a(t, s_{k+1}), s_k)] \geq 0. \quad (\text{O})$$

Proof of Lemma 1. Fix t . Write $S_t := \text{supp} p_t$ and $a_s := a(t, s)$.

Suppose first that r_t supports the policy. For each k , apply the supporting inequality at signal s_k to the feasible action $a_{s_{k+1}}$:

$$r_t(a_{s_k}) - r_t(a_{s_{k+1}}) \geq U_t(a_{s_{k+1}}, s_k) - U_t(a_{s_k}, s_k).$$

Summing over k cancels the rewards and gives

$$0 \geq \sum_{k=1}^K [U_t(a_{s_{k+1}}, s_k) - U_t(a_{s_k}, s_k)],$$

which is condition (O).

Conversely, suppose condition (O) holds. For $s, s' \in S_t$, define the deviation gain

$$w_t(s, s') := U_t(a_{s'}, s) - U_t(a_s, s).$$

Condition (O) is equivalent to

$$\sum_{k=1}^K w_t(s_k, s_{k+1}) \leq 0$$

for every directed signal cycle. Define

$$V_t(s) := \sup_{m \geq 0, \substack{s_0, \dots, s_m \in S_t \\ s_0 = s}} \sum_{j=0}^{m-1} w_t(s_j, s_{j+1}),$$

where the empty path has weight zero. Since S_t is finite and every cycle has nonpositive weight, deleting cycles from any path raises its weight. The supremum is thus finite and attained by a simple path.

For any path starting at s' , the edge $s \rightarrow s'$ followed by that path forms a path starting at s . Therefore,

$$V_t(s) \geq w_t(s, s') + V_t(s').$$

If $a_s = a_{s'}$, then $w_t(s, s') = w_t(s', s) = 0$. Applying this inequality in both directions gives $V_t(s) = V_t(s')$. Hence the following reward is well defined:

$$r_t(a_s) := V_t(s) - \max_{q \in S_t} V_t(q), \quad r_t(a') = -\infty \quad \text{if } a' \notin \{a_s : s \in S_t\}.$$

For any assigned action $a_{s'}$, the same inequality gives

$$\begin{aligned} U_t(a_s, s) + r_t(a_s) - U_t(a_{s'}, s) - r_t(a_{s'}) \\ = V_t(s) - V_t(s') - w_t(s, s') \geq 0. \end{aligned}$$

Unassigned actions are prohibited. Thus r_t supports the policy, and its largest on-path value is zero. $\square$

Fix one supporting schedule $r_{\hat{t}}$ for each prescribed policy $a(\hat{t}, \cdot)$. Type t 's interim value from the continuation schedule intended for type $\hat{t}$ is

$$G_t(\hat{t}) := \sum_{s \in \text{supp } p_t} p_t(s) \max_{a' \in A[\hat{t}]} \{U_t(a', s) + r_{\hat{t}}(a')\}.$$

The true signal law p_t appears after every report. A deviating type reoptimizes after its signal and need not follow type $\hat{t}$'s prescribed policy. Note here that only limited deviations are available: type t can report $\hat{t}$ only if $t \geq \hat{t}$.

Given these supporting schedules, the truthful-reporting condition requires every type cycle $t_1, \dots, t_K$, with $t_{K+1} = t_1$ and $t_k \geq t_{k+1}$ and $G_{t_k}(t_{k+1}) > -\infty$ for every k , to satisfy

$$\sum_{k=1}^K [G_{t_k}(t_k) - G_{t_k}(t_{k+1})] \geq 0. \quad (\text{T})$$

We say that a policy is implementable by a state-independent mechanism if the direct mechanism recommends $a(\hat{t}, \cdot)$ after report $\hat{t}$ and its reward satisfies $r(\hat{t}, a, \theta) = r(\hat{t}, a)$.

Proposition 2 (Cycle characterization). A physically feasible policy is implementable by a state-independent mechanism if and only if condition (O) holds for every type and supporting schedules can be chosen so that condition (T) holds.

Proof of Proposition 2. We prove necessity, then construct.

Suppose a state-independent schedule $\tilde{r}$ implements the policy. For each report t , let

$$B_t := \{a(t, s) : s \in \text{supp } p_t\}.$$

Set $\tilde{r}(t, a) = -\infty$ for $a \notin B_t$. This change leaves truthful utility unchanged and weakly lowers the value of every false report.

Define

$$c_t := \max_{a \in B_t} \tilde{r}(t, a), \quad r_t(a) := \tilde{r}(t, a) - c_t.$$

The schedule r_t is a normalized supporting schedule. Let W_t denote type t 's truthful interim utility under $\tilde{r}$. Since adding c_t shifts every continuation payoff after report t by the same amount,

$$W_t = G_t(t) + c_t, \quad \text{and hence} \quad c_t = W_t - G_t(t).$$

A true type t that reports $\hat{t}$ and then chooses its best feasible action obtains

$$G_t(\hat{t}) + c_{\hat{t}} = G_t(\hat{t}) + W_{\hat{t}} - G_{\hat{t}}(\hat{t}).$$

Incentive compatibility therefore requires

$$W_t - W_{\hat{t}} \geq G_t(\hat{t}) - G_{\hat{t}}(\hat{t})$$

whenever $t \geq \hat{t}$ and $G_t(\hat{t}) > -\infty$.

Now take a feasible type cycle $t_1, \dots, t_K$, with $t_{K+1} = t_1$. Summing these inequalities gives

$$\begin{aligned} 0 &= \sum_{k=1}^K (W_{t_k} - W_{t_{k+1}}) \\ &\geq \sum_{k=1}^K [G_{t_k}(t_{k+1}) - G_{t_{k+1}}(t_{k+1})] \\ &= - \sum_{k=1}^K [G_{t_k}(t_k) - G_{t_k}(t_{k+1})]. \end{aligned}$$

This is condition (T). Since each r_t supports the prescribed signal policy, [Lemma 1](#) also gives condition (O).

Conversely, fix normalized supporting schedules that satisfy condition (T). Put a directed edge $t \rightarrow \hat{t}$ whenever

$$t \geq \hat{t} \quad \text{and} \quad G_t(\hat{t}) > -\infty,$$

and give this edge weight

$$q(t, \hat{t}) := G_t(\hat{t}) - G_{\hat{t}}(\hat{t}).$$

For every directed type cycle,

$$\sum_{k=1}^K q(t_k, t_{k+1}) = - \sum_{k=1}^K [G_{t_k}(t_k) - G_{t_k}(t_{k+1})] \leq 0.$$

Define

$$W_t := \sup_{\substack{m \geq 0, t_0, \dots, t_m \in \mathbb{T} \\ t_0 = t, (t_j, t_{j+1}) \text{ is an edge}}} \sum_{j=0}^{m-1} q(t_j, t_{j+1}).$$

The empty path has weight zero. Since $\mathbb{T}$ is finite and every cycle has nonpositive weight, W_t is finite and attained by a simple path. Prepending an edge gives

$$W_t \geq q(t, \hat{t}) + W_{\hat{t}} = G_t(\hat{t}) - G_{\hat{t}}(\hat{t}) + W_{\hat{t}}.$$

Define the mechanism's reward by

$$r(t, a) := r_t(a) + W_t - G_t(t).$$

The shift does not change obedience. Truthful reporting gives type t the interim utility

$$G_t(t) + W_t - G_t(t) = W_t.$$

A feasible false report $\hat{t}$ gives

$$G_t(\hat{t}) + W_{\hat{t}} - G_{\hat{t}}(\hat{t}) = q(t, \hat{t}) + W_{\hat{t}} \leq W_t$$

by the path inequality above. Thus the schedule implements the policy. A common shift of all W_t changes no incentive constraint. $\square$

We make several remarks about **Proposition 2**. First, **Figure 2** gives a graphical interpretation of the two conditions. Within report t , panel (a) uses an arrow $s \rightarrow s'$ for the gain from taking after signal s the action assigned after s' . The relative rewards r_t support the prescribed use of the private signal, and condition (O) rules out a positive signal cycle. Across reports, W_t is type t 's truthful interim utility, and panel (b) gives the arrow $t \rightarrow \hat{t}$ weight $G_t(\hat{t}) - G_{\hat{t}}(\hat{t})$. Each report $\hat{t}$ generates the payoff profile

$$t \mapsto G_t(\hat{t}) + W_{\hat{t}} - G_{\hat{t}}(\hat{t}),$$

which touches W at $t = \hat{t}$ and lies weakly below it at every other type. Thus condition (T) makes W a supporting potential on the report graph. In the standard Euclidean screening model these profiles become Rochet's supporting hyperplanes; here the finite graph formulation requires neither an order nor a vector structure on types.

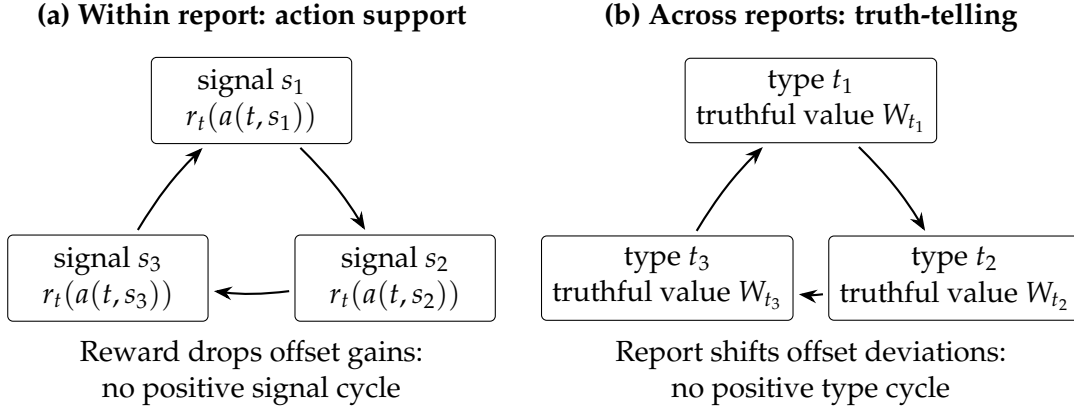


Figure 2: Nested supporting potentials.

Second, **Proposition 2** generalizes various notions of cyclical monotonicity within the screening literature. It recovers [Rochet \(1987\)](#) in the special case in which the agent receives no interim information, every type shares the same action set, and every type can report every other type. This is the top right cell of [Table 1](#). Define type t 's expected payoff from action a by

$$\bar{U}_t(a) := \sum_{\theta \in \Theta} h[t](\theta) u[t](a, \theta).$$

A policy now assigns one action $a(t)$ to each type. After report $\hat{t}$, prohibit every action except $a(\hat{t})$ and write $p_{\hat{t}}$ for the reward attached to that action. The agent therefore chooses only its report, and incentive compatibility becomes

$$\bar{U}_t(a(t)) + p_t \geq \bar{U}_t(a(\hat{t})) + p_{\hat{t}} \quad \text{for every } t, \hat{t}.$$

Condition **(O)** is vacuous, while condition **(T)** becomes Rochet's cycle condition:

$$\sum_{k=1}^K [\bar{U}_{t_k}(a(t_k)) - \bar{U}_{t_k}(a(t_{k+1}))] \geq 0 \quad \text{with } t_{K+1} = t_1.$$

Thus the outer potential W is Rochet's indirect-utility potential, with primitive payoffs replaced by expected payoffs. If types and actions are Euclidean and $\bar{U}_t(a) = t \cdot a$, the profiles that support W are the usual supporting hyperplanes of a convex function. If reports or feasible actions are restricted, condition **(T)** instead gives the same test on the resulting directed report graph.

	Type observed	Type private
No interim info	No nontrivial cycle. Rewards can support any physically feasible policy.	Rochet on expected utilities. Condition (T) is standard cyclical monotonicity with a common action set and complete report graph.
Interim info	Within-report cycles. Condition (O) screens the signal through the agent’s action choice.	Nested cycles. Condition (O) supports action choice after each report; condition (T) supports the initial report.

Table 1: Taxonomy of cyclical monotonicity.

The bottom-right cell is closest to [Krähmer and Strausz \(2017\)](#) that show an equivalence between sequential screening (with interim information) and static screening where the whole continuation contract serves as the allocation. Our outer problem uses a similar reduction.¹³ The difference lies in the continuation problem: [Krähmer and Strausz \(2017\)](#) use one-dimensional linear preferences, which reduce second-stage incentives to monotonicity and an envelope formula. We allow finite but unordered signals, actions, preferences, and feasible sets. Moreover, agents might undertake double deviations—a type that makes a false initial report might later disobey. We must thus keep the two support problem separate.

Third, for finite spaces the characterization gives an exact numerical procedure, though not a general polynomial-time solution to the designer’s problem. Fix a physically feasible deterministic policy. A linear feasibility program can decide in polynomial time whether the fixed policy is implementable.¹⁴ Optimizing over policies is more complicated and requires more structure.¹⁵ For example, one could order types, signals, and actions and impose single crossing strong enough to reduce both cycle conditions to monotonicity. The designer would then search over monotone policies rather than all policies.

¹³Treat the normalized reward schedule $r_{\hat{t}}$ as the continuation allocation, let type t value it at $G_t(\hat{t})$, and let the report-specific constant $c_{\hat{t}}$ act as its transfer. The initial incentive constraint is then

$$G_t(t) + c_t \geq G_t(\hat{t}) + c_{\hat{t}},$$

so condition (T) is static cyclical monotonicity over continuation reward schedules. Their second-stage incentive constraint plays the role of condition (O), while their first-stage constraint plays the role of condition (T).

¹⁴The variables are the mechanism’s on-path rewards, each type’s truthful interim utility, and an auxiliary continuation value for every feasible false report and signal. Obedience and truth-telling give linear inequalities.

¹⁵See, e.g., [Figalli et al. \(2011\)](#), who identify a nonnegative cross-curvature condition under which a class of multidimensional screening problems reduces to a convex program.

2.5 Dimensions of single-agent mechanisms. We now conclude with a brief taxonomy of mechanisms within our framework. A reward can respond only to what the evaluator can identify about the action and the state. Indeed, the difficulty of specifying a reward for an AI agent has been recognized as an important obstacle to alignment (Hadfield-Menell and Hadfield, 2019; Hadfield-Menell, 2021). As in an incomplete contract, a coarse measure leaves some payoff-relevant distinctions outside the reward schedule. Table 2 varies the precision of these two measurements while allowing unrestricted rewards and prohibitions.

	No state measure	Coarse state measure	Exact state measure
No action measure	<i>Futile.</i>	<i>Coarse state scoring.</i> Scores distinguish beliefs only through the measured state signal. Full elicitation of AI’s information.	<i>State scoring.</i> Scores can elicit the agent’s belief about θ , but not preference or capability differences. Full elicitation of AI’s information.
Coarse action measure	<i>Coarse action rewards.</i> Rewards distinguish actions only through what can be measured and reinforced. Koh (2026).	<i>Coarse performance rewards.</i> Rewards respond to the measured action and state signals.	<i>State-contingent coarse rewards.</i> Action rewards vary with the signal about state.
Exact action measure	<i>State-independent action rewards.</i> The designer can pick the reward schedule but same instrument applies in every state. Examples I, II and V.	<i>Coarsely contingent action rewards.</i> Designer can condition rewards precisely on actions but only on noisy state measures.	<i>Full contractability.</i> The designer can price every action–state pair. First-best.

Table 2

3 APPLICATIONS: SINGLE AGENT

We give two stylized applications of our framework. Payoffs are given by the quadratic loss function

$$v(a, \theta) = -(a - \theta)^2 \quad \text{and} \quad u_b(a, \theta) = -(a - \theta - b)^2,$$

where $\theta \in \Theta \subseteq \mathbb{R}$. Here, we might think of θ as capturing some fundamental state of the world that is external to both the human and AI. We want the AI’s actions to be *apt* for the world: if θ is the disease we want AI to pick the right treatment; if θ is a software flaw, we want AI to select the right patch, and so on. AI’s preferences are parametrized by the bias $b \geq 0$, and the common action space is $A = \mathbb{R}$. Throughout our applications, we suppose that the AI learns the state perfectly. Both applications use deterministic mechanisms.

We emphasize that these applications are deliberately stylized such as to make the basic forces transparent and legible to a non-technical reader. We have chosen the linear-quadratic framework because it is (i) *tractable*: best responses are linear and optimal caps can be described in simple closed forms; (ii) *apt*: we want to capture the idea that actions should be appropriate for the ‘state of the world’; (iii) *doing and knowing*: the ability for AI to undertake apt actions depend on their ability to both *do* (action space) and *know* (what they learn about the state); (iii) *geometry*: we have a notion of an action being more or less apt for the state as measured by the quadratic distance.

Our goal is to illustrate how our framework might map onto aspects of the alignment problem, to rationalize existing practices in AI safety, and to offer a common language of incentives and information through which alignment failures can be understood. Our first application ([Example I](#)) develops a simple analysis of sandbagging in which AI capabilities (size of the action set) and bias (b) are unknown. Our second application ([Example II](#)) focuses on a tradeoff between what we call ‘interpretability’: the dispersion of the designer’s beliefs about the AI’s bias; and ‘expected alignment’: the mean of the AI’s alignment $\mathbb{E}_\Gamma[b]$ under the designer’s prior.

Example I (Sandbagging). It is standard practice among the leading AI labs to specify a policy that conditions deployment on evaluated capability:

“The basic idea is to require safety, security, and operational standards appropriate to a model’s potential for catastrophic risk, with higher ASL [AI Safety Level] levels requiring increasingly strict demonstrations of safety.” —Anthropic’s *Responsible Scaling Policy* ([2024](#))

“[G]eneral availability deployment of a model takes place only after the appropriate corporate governance body determines the safety case regarding each CCL [Critical Capability Level] the model has reached to be adequate.” —Google DeepMind’s *Frontier Safety Framework* ([2025](#))

“We do not deploy models that reach a High capability threshold until the associated risks that they pose are sufficiently minimized.” —OpenAI’s *Preparedness Framework* ([2025b](#))

Such policies can be broadly construed as a mechanism in the sense we’ve defined: a public commitment mapping capability reports—e.g., evaluation scores—into deployment permissions. In this example we ask how such a map should be designed when the score is chosen strategically. Capability evaluations matter when scores change deployment: for instance, if a lower score buys more access, a more capable AI system is then incentivized to underperform.¹⁶ Frontier models also sandbag in context: when told that a high score triggers restriction or unlearning, models

¹⁶Models can underperform on dangerous-capability evaluations while they preserve general benchmark scores, and a password can hide or reveal their capabilities ([van der Weij et al., 2025](#)).

strategically underperform (Meinke et al., 2024; OpenAI, 2024). We call this strategic underperformance *sandbagging*.

Of course, sandbagging can only arise if models ‘anticipate’ that their performance in evaluations might shape future rewards e.g., deployment decisions. There is now some evidence for these kinds of strategic reasoning in ‘evaluation awareness’: Needham et al. (2025) find that frontier models distinguish evaluation from deployment transcripts well above chance. Likewise, Anthropic’s and OpenAI’s recent system cards indicate some evaluation awareness.¹⁷

Furthermore, our verification order will allow more capable types to imitate less capable types. This is fairly well-documented: concealing capabilities tends to be robust to countermeasures (Greenblatt et al., 2024b; Hofstätter et al., 2025). Models can also resist elicitation by reinforcement learning itself, strategically suppressing exploration so that the capability is never expressed during training (Jang et al., 2026).

Setup The human commits to a permission policy for a single AI agent whose capability and preferences are unknown. The AI’s capability c is drawn from an atomless distribution F with full support on $[\underline{c}, \bar{c}] \subseteq (0, 1]$. Type c has the action set

$$A(c) = (-\infty, c].$$

Its bias is the known continuous function $b(c)$, where

$$0 \leq b(c) \leq \min\{c, 1/2\}.$$

Note here that we have ‘projected’ our type space onto a single dimension since capabilities move with bias. The state $\theta \sim U[0, 1]$ drawn, the AI observes it, and the AI acts. The state is not contractible as in standard models of delegation (Holmström, 1979). Since types are pinned down by c , we henceforth use it to refer to the AI’s type. Our verification order combines execution and information capability:

$$c \geq \hat{c} \iff A(c) \supseteq A(\hat{c}) \text{ and } \pi[c] \text{ Blackwell dominates } \pi[\hat{c}].$$

All types observe the state perfectly in this example, so their experiments coincide. The order therefore reduces to $c \geq \hat{c}$ if and only if $c \geq \hat{c}$: a type can certify any lower capability but cannot certify a higher one.¹⁸

For simplicity, we will focus on delegation mechanisms that cap the action space from above.¹⁹ In particular, a report $\hat{c}$ generates the cap $\bar{a} : \hat{c} \rightarrow \mathbb{R}$ and so the permitted set

¹⁷A training snapshot related to Anthropic’s Mythos 5 became more aware of its grader during training; Opus 5 can distinguish evaluations from deployment when asked, even though it rarely states this belief in more realistic tests (Anthropic, 2026a,b). OpenAI’s GPT-5 System Card reports frequent explicit reasoning about the purpose of evaluations (OpenAI, 2025a).

¹⁸This is related to Gan and Li (2026) whose agent can verifiably disclose a subset of its privately known feasible set but it is the principal who picks the action.

¹⁹Such mechanisms are not always exactly optimal among the class of delegation mechanisms that take values in $\{0, -\infty\}$ but we think they form a natural class and will substantially simplify exposition. They are also ‘near’ optimal in simulations.

of actions is

$$D(\hat{c}) = (-\infty, \bar{a}(\hat{c})].$$

The threshold $\bar{a}(\hat{c})$ is the AI's highest permitted action; it can represent a spending, tool-use, or authorization limit.

A nominal cap $\bar{a}$ on type c induces the action

$$a(c, \theta) = \min\{\theta + b(c), c, \bar{a}\}.$$

where the first term is the AI's optimum if its in the interior; the second term is when the AI's capability binds—it would like to take a higher action but is simply unable to do so; and the last term is when the cap binds.

For a cap $\bar{a} \in [0, c]$, direct integration gives the human's expected utility

$$V_b(\bar{a}) := \mathbb{E}[v(\min\{\theta + b, \bar{a}\}, \theta)] = \begin{cases} -\bar{a}^2 + \bar{a} - \frac{1}{3}, & \bar{a} \leq b, \\ -\bar{a}b^2 - \frac{(1-\bar{a})^3}{3} + \frac{2b^3}{3}, & \bar{a} > b. \end{cases}$$

The expected utility is continuously differentiable and strictly concave, with

$$\frac{\partial V_b(\bar{a})}{\partial \bar{a}} = \begin{cases} 1 - 2\bar{a}, & \bar{a} \leq b, \\ (1 - \bar{a})^2 - b^2, & \bar{a} > b. \end{cases}$$

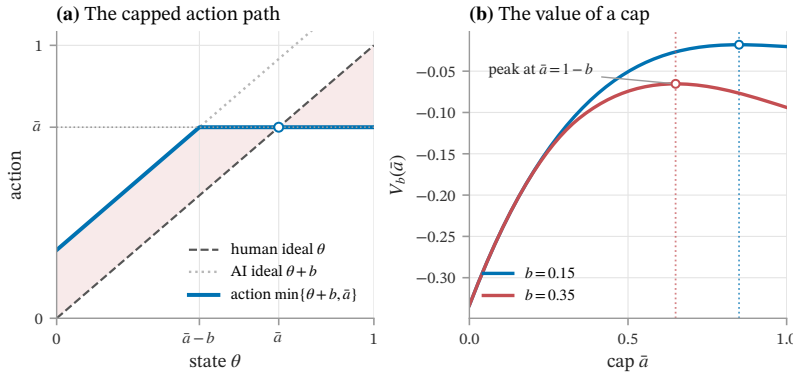


Figure 3: The anatomy of capped action.

Panel (a): with bias $b = 0.25$ and cap $\bar{a} = 0.7$ the AI tracks its ideal $\theta + b$ until the cap binds at $\theta = \bar{a} - b$. Panel (b): human's expected utility $V_b(\bar{a})$ is strictly concave and peaks at $\bar{a} = 1 - b$, so more biased types warrant tighter caps

Now notice that type c can pretend to be any type $\hat{c} \leq c$ but not $\hat{c} > c$. Hence, passing a harder test delivers a lower bound on capability but not vice versa.²⁰

²⁰Although recent techniques have tried to elicit model capabilities e.g., by injecting a small amount of noise (Tice et al., 2024) and doing so might further refine the verification order by also deterring certain kinds of sandbagging.

We can normalize the cap after score $\hat{c}$ so that $\bar{a}(\hat{c}) \leq \hat{c}$ since lowering a cap to the reported frontier leaves that reporter's action unchanged and weakly reduces the gain from sandbagging. Hence the threshold $\bar{a}(c)$ is the effective cap under truthful reporting.

Now observe that the schedule $\bar{a}(c) \leq c$ is truthfully implementable if and only if it is weakly increasing. If it is increasing, every lower certificate gives type c a weakly smaller permitted set. If $\bar{a}(\hat{c}) > \bar{a}(c)$ for some $\hat{c} < c$, type c can instead certify $\hat{c}$ and use the looser cap. This monotonicity condition is in the spirit of condition (T). Our problem thus reduces to the convex problem

$$\begin{aligned} & \max_{\bar{a}(\cdot)} \int V_{b(c)}(\bar{a}(c)) \, dF(c) \\ & \text{subject to } 0 \leq \bar{a}(c) \leq c \quad \text{and} \quad \bar{a}(\cdot) \text{ weakly increasing.} \end{aligned}$$

We did not impose any structure on how bias varies with capabilities so the question is whether the cap chosen *as if* capabilities are known already rises with capability. If so, then we are done; if not, then the monotonicity constraint binds and we have to iron. Here, the procedure is a little non-standard but nonetheless fairly straightforward: let $z^*(c)$ solve the auxiliary problem

$$\begin{aligned} z^* \in \arg \min_{z(\cdot)} & \int [z(c) - b(c)]^2 \, dF(c) \\ & \text{subject to } z(\cdot) \text{ measurable and weakly decreasing,} \\ & (1 - c)^2 \leq z(c) \leq 1 \quad \text{for all } c \in [\underline{c}, \bar{c}]. \end{aligned} \tag{IRON}$$

Thus z^* is the decreasing least-squares fit of squared bias, subject to the physical capability bound.

Proposition 3 (Optimal caps). Problem (IRON) has a solution z^* that is unique up to F -null sets. The corresponding optimal cap is, up to F -null sets,

$$\bar{a}^*(c) = 1 - \sqrt{z^*(c)}.$$

It attains the complete-information payoff if and only if $\min\{c, 1 - b(c)\}$ is weakly increasing. In particular:

(i) If $b(c)$ is weakly decreasing, the optimal evaluated cap is

$$\bar{a}^*(c) = \min\{c, 1 - b(c)\},$$

and evaluation attains the complete-information payoff.

(ii) If $b(c)$ is weakly increasing, evaluation and the optimal common cap give the same payoff. The optimal cap is

$$\bar{a}^*(c) = \min\{c, a^*\},$$

where

$$a^* = \begin{cases} \bar{c}, & \text{if } \bar{c} \leq 1 - b(\bar{c}), \\ 1 - \sqrt{\mathbb{E}[b(c)^2 \mid c > a^*]}, & \text{if } \bar{c} > 1 - b(\bar{c}). \end{cases}$$

The second branch has a unique solution in $[1/2, \bar{c})$.

Proof of Proposition 3. We first derive the ironing formula, then establish the two special cases.

Step 1: ironing. Replacing any feasible cap schedule $\bar{a}(c)$ by

$$\max\{\bar{a}(c), \min\{c, 1/2\}\}$$

preserves feasibility. It also weakly raises the human's payoff, and raises it strictly wherever the cap changes. Indeed, $V'_b(\bar{a}) > 0$ for $\bar{a} < \min\{c, 1/2\}$: if $\bar{a} \leq b$, then $V'_b(\bar{a}) = 1 - 2\bar{a} > 0$, while if $\bar{a} > b$, then

$$V'_b(\bar{a}) = (1 - \bar{a})^2 - b^2 > (1 - \bar{a})^2 - \bar{a}^2 = 1 - 2\bar{a} > 0.$$

An optimal cap therefore satisfies $\bar{a}(c) \geq \min\{c, 1/2\} \geq b(c)$. On this range,

$$V_{b(c)}(\bar{a}(c)) = -\bar{a}(c)b(c)^2 - \frac{(1 - \bar{a}(c))^3}{3} + \frac{2b(c)^3}{3}.$$

Set $z(c) = (1 - \bar{a}(c))^2$. A cap schedule is weakly increasing and satisfies $0 \leq \bar{a}(c) \leq c$ if and only if $z(\cdot)$ is weakly decreasing and satisfies $(1 - c)^2 \leq z(c) \leq 1$. After dropping terms that do not depend on the cap, the negative of expected utility becomes

$$\int \left[\frac{z(c)^{3/2}}{3} - b(c)^2 \sqrt{z(c)} \right] dF(c).$$

The following lemma connects the least-squares problem to the transformed loss.

Lemma 2 (Continuum ironing). Problem (IRON) has a pointwise weakly decreasing solution z^* that satisfies

$$(1 - c)^2 \leq z^*(c) \leq (1 - \min\{c, 1/2\})^2 \quad \text{for every } c.$$

Any other solution is equal to z^* F -almost surely. The same function z^* also minimizes

$$\int \left[\frac{z(c)^{3/2}}{3} - b(c)^2 \sqrt{z(c)} \right] dF(c)$$

over all pointwise weakly decreasing functions satisfying these bounds. Every minimizer of this second problem is equal to z^* F -almost surely.

The proof is in [Appendix A](#). We therefore recover the optimal cap from $\bar{a}^*(c) = 1 - \sqrt{z^*(c)}$.

With known capability, strict concavity gives the unique pointwise cap $\min\{c, 1 - b(c)\}$. Evaluation attains this bound exactly when that cap is weakly increasing and hence truthfully implementable. If $b(c)$ is weakly decreasing, both c and $1 - b(c)$ are weakly increasing. Their minimum is therefore weakly increasing, which proves part (i).

Step 2: the common cap. Let

$$a^* \in \arg \max_{\bar{a} \in [0, \bar{c}]} \int V_{b(c)}(\min\{c, \bar{a}\}) dF(c).$$

The objective is continuous on a compact interval, so a maximizer exists. Any cap below $1/2$ that binds a positive mass is not optimal. For an affected type with $\bar{a} \leq b$, we have $V'_b(\bar{a}) = 1 - 2\bar{a} > 0$. For one with $b < \bar{a} < 1/2$,

$$V'_b(\bar{a}) = (1 - \bar{a})^2 - b^2 > (1 - \bar{a})^2 - \bar{a}^2 = 1 - 2\bar{a} > 0.$$

Raising the cap therefore raises every affected type's utility. Any binding optimum thus obeys $a^* \geq 1/2 \geq b(c)$. Because F is atomless, only types with $c > \bar{a}$ change their actions after a marginal cap change. At a binding optimum,

$$\int_{\{c > a^*\}} \left[(1 - a^*)^2 - b(c)^2 \right] dF(c) = 0.$$

Dividing by $\mathbb{P}(c > a^*)$ and taking the nonnegative square root gives the conditional formula. If every type can reach the cap, the conditional expectation is unconditional. When b is weakly increasing, the conditional expectation is weakly increasing in the

threshold. The right side of the fixed-point equation is therefore weakly decreasing, so an interior solution is unique.

Step 3: increasing bias. Suppose b is weakly increasing and let $\bar{a}^*(c) = \min\{c, a^*\}$. This schedule is weakly increasing and bounded above by c , so it is feasible. Consider first a binding common cap, $a^* < \bar{c}$, and any other feasible schedule $\bar{a}$. Concavity gives

$$\int \left[V_{b(c)}(\bar{a}(c)) - V_{b(c)}(\bar{a}^*(c)) \right] dF(c) \leq \int V'_{b(c)}(\bar{a}^*(c)) [\bar{a}(c) - \bar{a}^*(c)] dF(c).$$

On the region $c \leq a^*$, the first factor is nonnegative and the second is nonpositive. Indeed, feasibility gives $\bar{a}(c) - c \leq 0$. If $V'_{b(c)}(c) < 0$ for one such type, then $b(c) > 1 - c \geq 1 - a^*$. Increasing bias would make

$$(1 - a^*)^2 - b(c')^2 < 0$$

for every $c' > a^*$, contrary to the optimal-cap condition above. Hence the product is nonpositive.

On the region $c > a^*$, the functions

$$(1 - a^*)^2 - b(c)^2 \quad \text{and} \quad \bar{a}(c) - a^*$$

are weakly decreasing and weakly increasing, respectively. The first has conditional mean zero by the optimal-cap condition. Chebyshev's integral inequality therefore makes their product integrate to a nonpositive value. No feasible evaluated schedule improves on $\bar{a}^*$. Strict concavity gives uniqueness up to F -null sets.

It remains to characterize the corner. If $a^* = \bar{c}$ and $c > 1 - b(c)$ for some type, continuity and increasing bias give a positive upper interval on which a small reduction in the common cap raises utility. This contradicts optimality at $\bar{c}$. Conversely, if $\bar{c} \leq 1 - b(\bar{c})$, increasing bias implies $c \leq 1 - b(c)$ for every type. Full access is then pointwise optimal, so $a^* = \bar{c}$. If the inequality fails, the optimum is interior and Step 2 gives the unique fixed point in part (ii). This proves the claim. $\square$

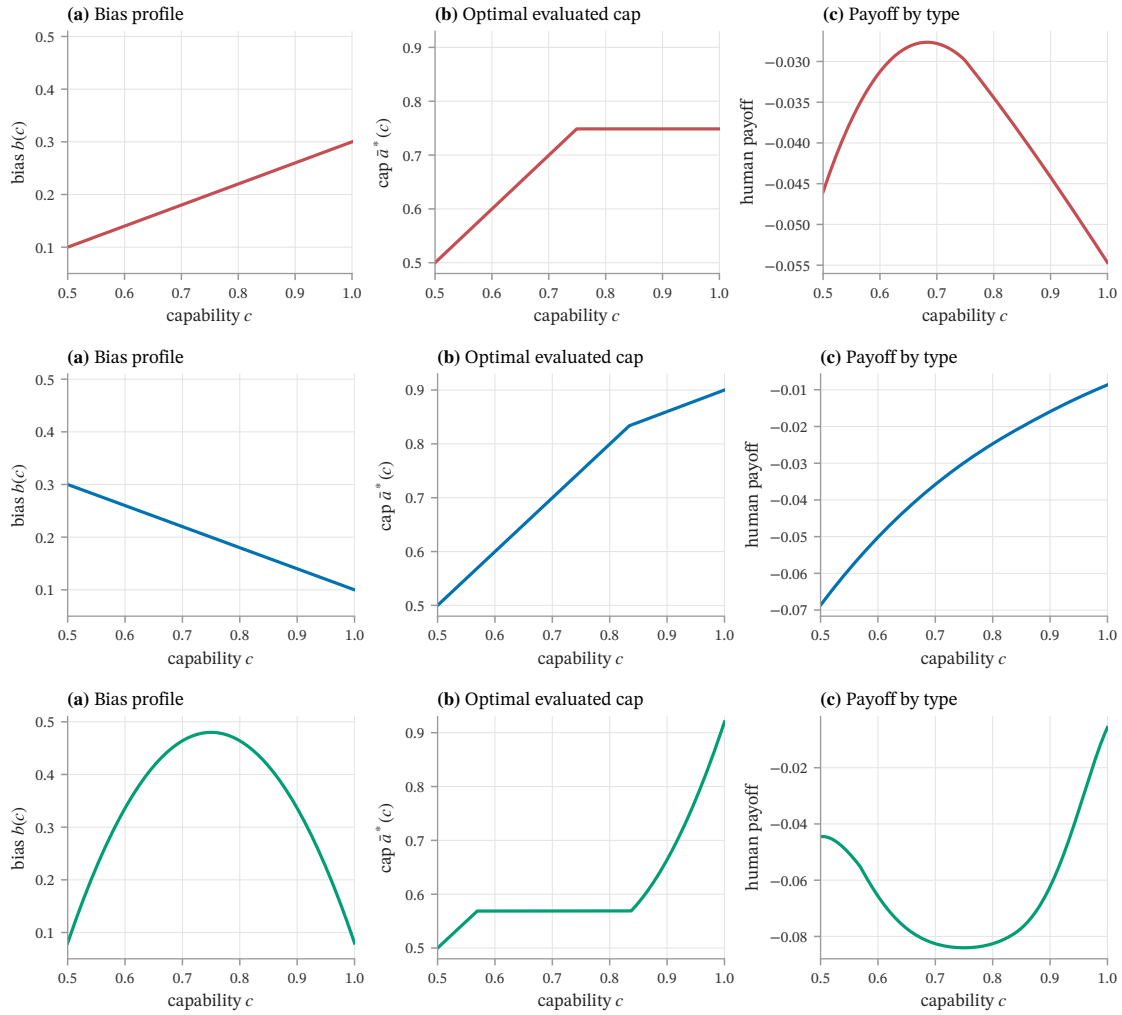


Figure 4: Optimal evaluated caps across bias profiles.

Top to bottom: rows use increasing bias $b(c) = 0.4c - 0.1$, decreasing bias $b(c) = 0.5 - 0.4c$, and nonmonotone bias $b(c) = 0.08 + 1.6x(1 - x)$, where $x = 2c - 1$. The columns show bias, the optimal evaluated cap, and the human's payoff by type.

Viewed through our model, sandbagging is profitable only when pretending to be a less capable type generates a looser cap. We might think of three cases. First, if bias rises with capability, powerful models are also more misaligned. The complete-information cap can then fall with capability. Such a schedule is not truthfully implementable: a more capable AI can pretend to be less capable and receive a looser cap. The implemented cap must instead rise weakly with capability, so the best feasible policy pools from above where this constraint binds. This is illustrated by the first row of [Figure 4](#).

Now suppose, instead, that bias falls with capability—as in the idea that more capable systems might converge to human values. The complete-information cap is then

weakly increasing, so a more capable AI cannot obtain a looser cap by pretending to be less capable. A less capable but more biased AI cannot pretend to be more capable—this is the sense in which capabilities cannot be faked. Hence, it is optimal for the designer to implement the cap *as if* it knew the AI’s capability. This is illustrated by the second row of [Figure 4](#).

Finally consider the case in which bias is nonmonotone in capability e.g., although a very powerful AI system might be aligned, an intermediate system might be quite biased. [Proposition 3](#) characterizes how the cap is optimally ironed such that intermediate types are pooled and face the same cap while relaxing the cap for extremely capable types who can be verified to be capable and hence aligned. ♦

Example II (The alignment–interpretability trade-off). Part of AI safety aims to interpret and understand the behavior of AI systems. What is the value of doing so? We argue that the right measure must take into account optimal mechanisms.

Setup. We remain in the linear-quadratic environment. To focus on interpretability and alignment, we will assume that all AI types are maximally capable i.e., the action space is simply the real line for all types, and that all AI types learn the state perfectly. Thus, an AI type is pinned down by its bias $b \geq 0$. A *technology* is a distribution $\tau \in \Delta(\mathbb{R})$ whose support contains only nonnegative AI biases. We think of a technology as capturing the designer’s subjective uncertainty about the AI’s bias after some training process.²¹ For instance, methods such as reinforcement learning from human feedback might push the AI’s preferences toward that of the human, but there is substantial residual uncertainty about the preference that emerges at deployment ([Casper et al., 2023](#)). As before, the AI (acts as if it) knows its bias $b \geq 0$ before it is evaluated. After it is deployed, the AI agent learns the true state $\theta \in U[0, 1]$ and acts.

We study state-independent deterministic delegation. A report selects a nonempty closed set of permitted actions $D \subseteq A$.²² A *report-independent* permission set $D = (-\infty, \bar{a}]$ induces the action

$$\operatorname{argmax}_{a \leq \bar{a}} u_b(a, \theta) = \min\{\bar{a}, \theta + b\}.$$

Since every AI type can take the full set of actions A and observes the state perfectly, we assume that the verification order is complete:

$$C(b) = \operatorname{supp}(\tau) \quad \text{for every } b \in \operatorname{supp}(\tau).$$

Thus every type can submit every report. A direct delegation mechanism assigns a nonempty closed permission set D_b to each report b . The mechanism is incentive

²¹See [Herrmann et al. \(2026\)](#) who makes a similar point but in the context of whether to deploy an agent i.e., comparing ‘no AI’ vs ‘full discretion’.

²²Note that this is a somewhat larger class than [Example I](#) that forced the delegation set to be an interval with the same floor.

compatible if each type weakly prefers the set assigned to its truthful report to the set assigned to every other report. We restrict attention to mechanisms whose induced losses are measurable.

Technology frontier and alignment-interpretability tradeoffs. For the technology τ , we say that *mean alignment* is

$$\bar{b} := \mathbb{E}_{b \sim \tau}[b]$$

and its *interpretability* is

$$\sigma^2 := \mathbb{E}_{b \sim \tau}[(b - \bar{b})^2].$$

Mean alignment measures how biased an AI system is on average. By contrast, interpretability is a measure of dispersion of our subjective uncertainty as to whether the AI system is biased. Indeed, the practice of behavioral evaluations (Perez et al., 2023) and mechanistic interpretability—probes and circuit-level analyses of model internals (Sharkey et al., 2025)—sharpen our subjective uncertainty about the bias and, in this regard, can be broadly viewed as aiding interpretability.

Under full discretion—in which the AI’s permitted set is the full action space—the AI with bias b chooses its ideal action $\theta + b$. Hence, the expected loss can be decomposed as

$$\mathbb{E}_{b \sim \tau}[b^2] = \bar{b}^2 + \sigma^2$$

which is analogous to the bias-variance decomposition of an estimator’s squared error. On this view, (squared) mean alignment and interpretability trade off against each other one-for-one. But this is under full discretion; in what follows, we will characterize the optimal mechanism and, in so doing, pin down the value of technologies that are more aligned on average versus more interpretable.

The *technology frontier* is a family of bias distributions $(\tau_\kappa)_{\kappa \in [0,1]}$. We use κ to index training processes that lower mean bias but raise residual uncertainty. When $\tau = \tau_\kappa$, the mean bias $\bar{b}$ falls with κ while the variance σ^2 rises. Thus, a higher κ makes the AI less biased on average but leaves the designer less certain about its residual bias. This reduced-form trade-off covers training that improves average behavior while making the remaining failure modes harder to predict (Kirk et al., 2024; Baker et al., 2025; Korbak et al., 2025).

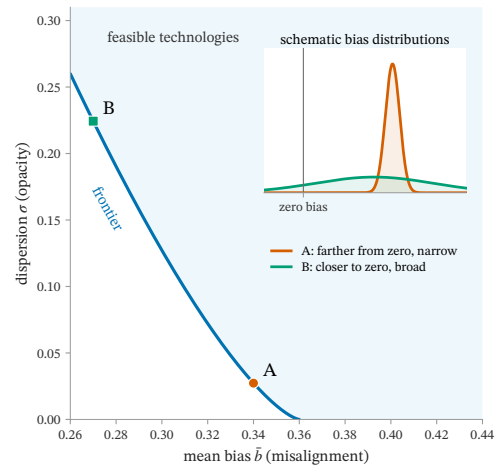


Figure 5: The technology frontier.

Optimal delegation. Fix a technology τ . What permission set should the human use? We will establish this in closed form under the condition

$$b \leq 1 - \sqrt{\bar{b}^2 + \sigma^2} \quad \tau\text{-almost surely.} \quad (\text{INT})$$

(INT) ensures that the optimal cap lies weakly above every realized bias, which gives the closed form below.

Proposition 4 (One cap is optimal). If (INT) holds, then

(i) the delegation set

$$D^* = (-\infty, \bar{a}^*] \quad \text{where} \quad \bar{a}^* = 1 - \sqrt{\bar{b}^2 + \sigma^2}$$

minimizes the human's loss;

(ii) the resulting minimum loss is

$$L^* = (\bar{b}^2 + \sigma^2) - \frac{2}{3}(\bar{b}^2 + \sigma^2)^{3/2} - \frac{2}{3}\mathbb{E}_{b \sim \tau}[b^3].$$

Proposition 4 states that a single cap at $\bar{a}^*$ performs at least as well as any incentive-compatible menu of delegation sets. The proof is somewhat involved and is in [Appendix A](#) though the basic idea is simple. Start from an arbitrary incentive-compatible mechanism. For each reported bias, we match the original delegation set to a report-dependent cap that gives the type the same utility as under the original mechanism. By construction, the each type remains indifferent. We then show that this matched cap weakly lowers the human's loss. Because every report is feasible for every type, incentive compatibility also implies that the endpoints of these matched caps rise weakly with reported bias. But the human prefers the opposite pattern: it prefers a tighter cap as bias rises. Thus, there is a single common cap that lies between the replaced cap and the human's optimal cap. Moving each type to this common cap weakly lowers the human's loss and remains incentive compatible.

We can then optimize over the common cap. The tradeoff here is standard in delegation ([Holmström, 1979](#); [Amador et al., 2006](#)): loosening the cap helps the AI track the state, while tightening it limits the effect of the AI's bias. Under condition (INT), the optimal balance sets the squared gap between the cap and the highest state equal to average squared bias.²³

²³The force is the same as in the increasing-bias case of [Proposition 3](#): a higher type can imitate any lower type while the human's preferred cap falls with bias. Here all types share the same action space, so pooling gives the same cap. Conversely, in [Example I](#) where capabilities were heterogeneous, the common cap was instead truncated by each type's capability.

The value of interpretability. To understand the value of interpretability, we consider a symmetric family of distributions

$$b = \bar{b} + \sigma\varepsilon,$$

where we fix a symmetric, finitely supported shock ε with $\mathbb{E}[\varepsilon] = 0$ and $\mathbb{E}[\varepsilon^2] = 1$. We restrict σ so that every realized bias is nonnegative and satisfies condition (INT). Symmetry gives $\mathbb{E}[\varepsilon^3] = 0$, so the centered third moment vanishes and we can write $\mathbb{E}[b^3] = \bar{b}^3 + 3\bar{b}\sigma^2$. Hence, the loss under the optimal mechanism is

$$L^*(\bar{b}, \sigma^2) = \bar{b}^2 + \sigma^2 - \frac{2}{3}(\bar{b}^2 + \sigma^2)^{3/2} - \frac{2}{3}(\bar{b}^3 + 3\bar{b}\sigma^2),$$

and the marginal effect of variance on loss is

$$\frac{\partial L^*}{\partial \sigma^2} = 1 - \sqrt{\bar{b}^2 + \sigma^2} - 2\bar{b} = \bar{a}^* - 2\bar{b}.$$

The basic idea is that a symmetric increase in variance moves residual bias away from its mean in both directions. Lower-bias realizations retain useful discretion, while the cap constrains higher-bias realizations. Because greater interpretability means lower variance, it lowers loss locally when $\bar{a}^* > 2\bar{b}$ and raises loss locally when $\bar{a}^* < 2\bar{b}$. The middle panel of [Figure 6](#) plots this derivative for a distribution of bias with binary support. When the threshold lies in the maintained range, its sign changes at $\sigma_{\dagger}^2 = (1 - \bar{b})(1 - 3\bar{b})$. The left panel of [Figure 6](#) compares the human's payoff under the optimal cap, full discretion, and no AI for a distribution of bias with binary support. Note that increasing σ can help because the cap limits the high-bias realization while preserving the low-bias realization. The right panel shows the same force through the value of delegation, which rises with residual uncertainty.

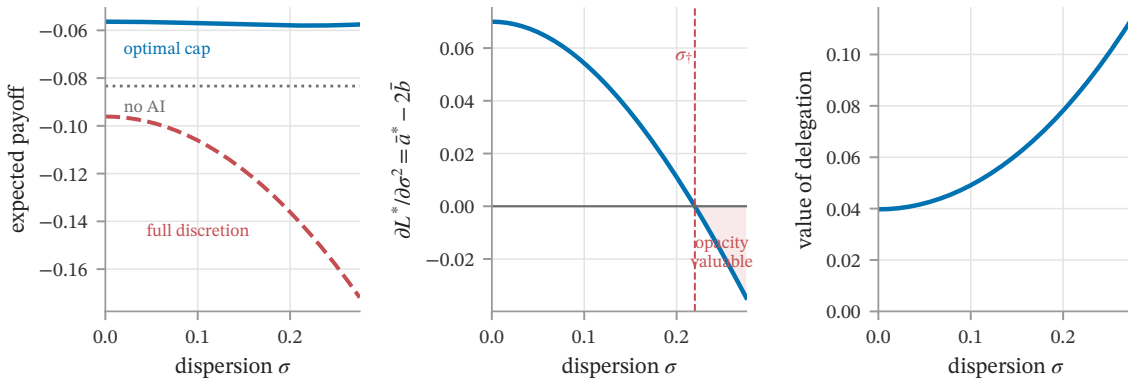


Figure 6: The value of interpretability at fixed mean bias.

The technology puts equal probability on $\bar{b} - \sigma$ and $\bar{b} + \sigma$, with $\bar{b} = 0.31$ and $\sigma_{\dagger} \approx 0.22$.

Moving along the frontier. The mechanism also changes which technology the designer wants to build. Under full discretion, loss is $\bar{b}^2 + \sigma^2$, so squared mean bias and variance trade one-for-one. Under the optimal cap, the third moment changes this ranking. For the parametrized symmetric family of distributions above,

$$\frac{\partial L^*}{\partial \bar{b}} = 2\bar{b} \bar{a}^* - 2(\bar{b}^2 + \sigma^2).$$

To illustrate, consider a frontier on which a higher κ buys lower mean bias at an increasing cost in residual uncertainty. **Figure 7** places this frontier on a loss surface. The left panel shows optimized loss. The red dashed locus is where the marginal value of interpretability changes sign. The filled dot is the technology chosen under the optimal cap; the open dot is the choice under full discretion i.e., without an optimal mechanism. Notice that the optimal cap truncates the loss from high bias realizations which in turn allows the designer to accept more residual uncertainty. The right panel shows the value of delegation.

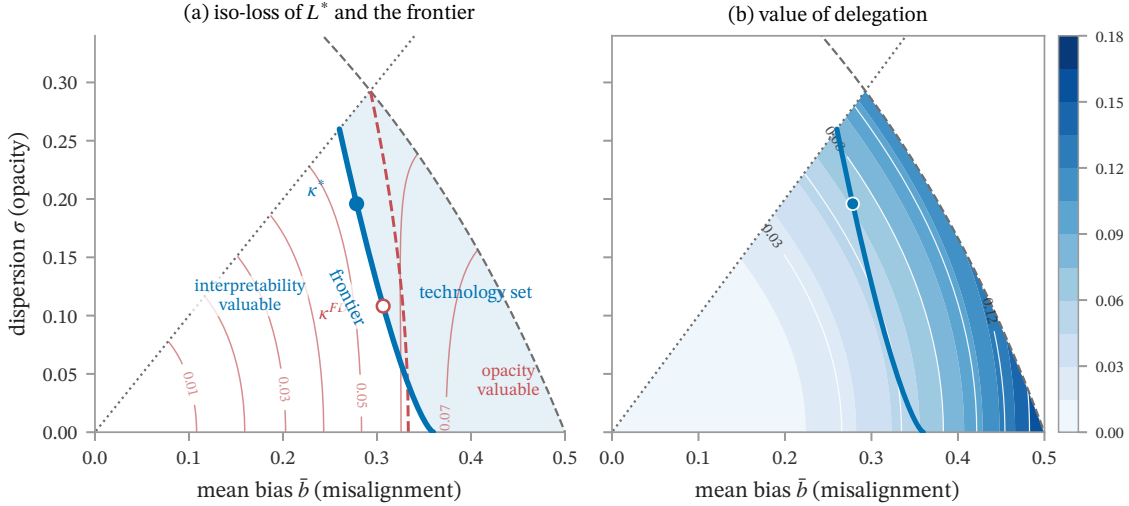


Figure 7: The technology frontier under optimal control.

Both panels use the two-point family and the frontier $\bar{b}(\kappa) = 0.36 - 0.10\kappa$ and $\sigma(\kappa) = 0.26\kappa^{1.4}$ for $\kappa \in [0, 1]$. Biases are nonnegative below the dotted line, and condition (INT) holds below the dashed grey curve.

Figure 8 then traces the mechanism and payoffs as the technology traverses the frontier. Note that the value of delegation changes along the frontier: the optimal delegation mechanism matters most when the technology leaves the AI either predictably biased or highly uncertain.

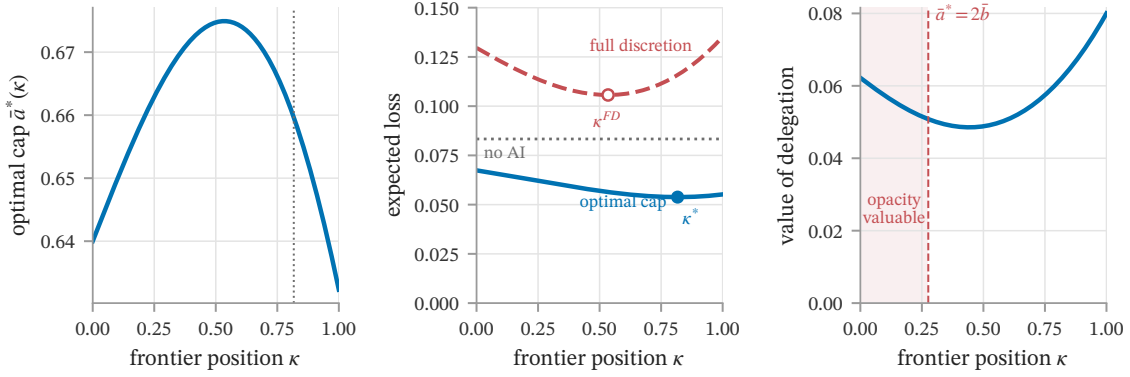


Figure 8: Moving along the technology frontier. The optimal cap first loosens as mean bias falls and then tightens as uncertainty grows. The full-discretion designer chooses $\kappa^{FD} \approx 0.53$, while the designer equipped with the optimal cap chooses $\kappa^* \approx 0.82$. The right panel shows the value of delegation along the frontier.

Proposition 5 (Substitutes in control, complements in value). Consider the parametrized family of symmetric distributions above.

- (i) *Substitutes in control.* The optimal cap depends on the technology only through $\bar{b}^2 + \sigma^2$. Hence, squared mean bias and variance are perfect substitutes in the chosen level of control.
- (ii) *Complements in value.* At every interior admissible parameter pair, the optimized loss satisfies

$$\frac{\partial^2 L^*}{\partial \bar{b} \partial \sigma^2} = -2 - \frac{\bar{b}}{\sqrt{\bar{b}^2 + \sigma^2}} < 0,$$

so, on every rectangle contained in the admissible parameter region, the optimized payoff $-L^*$ has increasing differences in alignment $-\bar{b}$ and interpretability $-\sigma^2$. Improving either raises the marginal value of the other.

Proof of Proposition 5. Part (i) follows from Proposition 4(i). For part (ii), the fixed distribution of ε gives

$$\mathbb{E}[b^3] = \bar{b}^3 + 3\bar{b}\sigma^2 + \sigma^3\mathbb{E}[\varepsilon^3].$$

The last term does not depend on $\bar{b}$. Differentiating the loss in Proposition 4(ii) therefore gives

$$\frac{\partial L^*}{\partial \bar{b}} = 2\bar{b}\bar{a}^* - 2(\bar{b}^2 + \sigma^2).$$

Since

$$\frac{\partial \bar{a}^*}{\partial \sigma^2} = -\frac{1}{2\sqrt{\bar{b}^2 + \sigma^2}},$$

we obtain

$$\frac{\partial^2 L^*}{\partial \bar{b} \partial \sigma^2} = -\frac{\bar{b}}{\sqrt{\bar{b}^2 + \sigma^2}} - 2.$$

Changing both arguments from $(\bar{b}, \sigma^2)$ to $(-\bar{b}, -\sigma^2)$ gives the increasing-differences claim. $\square$

Under full discretion, loss is $\bar{b}^2 + \sigma^2$, so the cross-partial is zero. Optimal control creates complementarity. A better-aligned technology receives more discretion, which gives interpretability more discretion to protect. At the same time, a more interpretable technology makes that discretion safer. Thus, our results formalize the sense in which understanding models is valuable—they give us, humans, confidence about the behavior of AI systems e.g., by probing language models (Goldowsky-Dill et al., 2025) and this, in turn, shapes the value of the optimal mechanism. $\blacklozenge$

4 MULTI-AGENT

We now extend our framework to multi-agent interactions. We will do so at quite a high level of generality, allowing such agents to hold beliefs about each other’s capabilities and preferences, beliefs about those beliefs, and so on. We think this model is valuable because it serves as a natural ‘limiting case’ for multi-agent rationality. Developing this machinery will allow our theory of multi-agent interaction to scale with agent capabilities.

While we think this machinery is valuable, we also want mechanisms that do not turn on the fine detail of agents’ higher-order beliefs (Bergemann and Morris, 2005), or for mechanisms to not demand too much of agents’ ability to optimize or reason strategically. Indeed, our multi-agent applications to follow will not rely very much on agents’ higher order beliefs. Nonetheless, we think that future notions of simplicity and robustness for AI agents are best understood in relation to this benchmark.

4.1 Environment and universal types. There are n agents indexed by $\mathcal{N} := \{1, \dots, n\}$. They choose an action profile $\mathbf{a} = (a_1, \dots, a_n) \in A^{\mathcal{N}}$. The human has payoff

$$v : A^{\mathcal{N}} \times \Theta \rightarrow \mathbb{R}.$$

Each agent privately knows its feasible action set and payoff. It also has beliefs about the state, the other agents, and their beliefs.

We will define objects to model agents' alignment and capabilities analogously to the single agent case. For each agent i , let

$$\mathcal{U}_i := \{f : A^{\mathcal{N}} \times \Theta \rightarrow \mathbb{R}\}.$$

Let H_i^0 be a singleton and define finite-order belief spaces recursively by

$$H_i^{k+1} := \Delta \left(\Theta \times \prod_{j \neq i} \underbrace{(\mathcal{A} \times \mathcal{U}_j)}_{\substack{j\text{'s available} \\ \text{actions \&} \\ \text{payoffs}}} \times H_j^k \right).$$

A hierarchy is *coherent* if every lower-order belief coincides with the relevant marginal of each higher-order belief; we restrict attention to coherent hierarchies. Following [Brandenburger and Dekel \(1993\)](#), whose construction covers our non-compact Polish component spaces (cf. [Mertens and Zamir, 1985](#)), we write a type as

$$t_i = (A_i[t_i], u_i[t_i], h_i[t_i]) \in \mathbb{T}_i,$$

where $h_i[t_i]$ is its coherent hierarchy. The universal type space admits a canonical belief map

$$t_i \longmapsto \beta_i[t_i] \in \Delta(\Theta \times \mathbb{T}_{-i}) \quad \text{where } \mathbb{T}_{-i} := \prod_{j \neq i} \mathbb{T}_j.$$

The belief $\beta_i[t_i]$ is the multi-agent counterpart of the belief $h[t] \in \Delta(\Theta)$ in the single-agent model. Its marginal on Θ is agent i 's belief about the state, while its joint law is the agent's prior over the state and other agents' types.

Write $\mathbb{T} := \prod_i \mathbb{T}_i$. The designer evaluates outcomes under a belief

$$\Gamma \in \Delta(\Theta \times \mathbb{T}).$$

For each agent i , let S_i be its interim signal space. An information structure is the map

$$\pi : \Theta \times \mathbb{T} \longrightarrow \Delta \left(\prod_i S_i \right),$$

where $s = (s_1, \dots, s_n)$ is the signal profile and agent i observes only s_i . Information structures can (i) correlate signals across agents; and (ii) signals can be informative of both the state $\theta \in \Theta$ as well as other players' types.

Under the designer's belief, the joint law of states, types, and signals is

$$\Gamma(d\theta, dt) \pi(ds \mid \theta, t) \in \Delta(\Theta \times \mathbb{T} \times S^{\mathcal{N}}).$$

Type t_i evaluates her strategies under the belief

$$\beta_i[t_i](d\theta, dt_{-i}) \pi(ds \mid \theta, t_i, t_{-i}) \in \Delta(\Theta \times \mathbb{T}_{-i}).$$

Thus agents and the designer agree on the information structure π but may disagree about the underlying state and types i.e., we need not have a common prior. But we will assume that, for every agent i , the conditional law $\Gamma(\cdot \mid t_i)$ of (θ, t_{-i}) is absolutely continuous with respect to $\beta_i[t_i]$ for almost every own type under the i -type marginal of Γ . This implies that agents do not rule out events that the designer considers possible.

Our representation of interim information π follows the standard decomposition of a Bayesian game into payoff-relevant uncertainty and an information structure (see, e.e., [Bergemann and Morris \(2016, 2019\)](#)). In our model, (θ, t) is the payoff-relevant uncertainty and the signals arrive between reports and actions. The single-agent experiment is the special case

$$\pi(ds \mid \theta, t) = \pi[t](ds \mid \theta).$$

Finally, we will once again allow for the possibility that agent reports (which can be interpreted as evaluations) are partially verifiable. As in the single-agent case, there is the reflexive and transitive order $\succeq$ ²⁴ and we write

$$C(t_i) := \{\hat{t}_i \in \mathbb{T}_i : t_i \succeq \hat{t}_i\}$$

for the set of types that $t_i \in \mathbb{T}_i$ can claim to be.

4.2 Multi-agent mechanisms.

Definition 3 (Multi-agent indirect mechanism). A multi-agent indirect mechanism is a tuple $\phi = (X, Y, \rho, R)$. For each agent i , the Borel spaces X_i and Y_i are its input and output alphabets. Write $X := \prod_i X_i$ and $Y := \prod_i Y_i$. The communication rule is

$$\rho : X \rightarrow \Delta(Y),$$

and the reward schedules $R = (R_i)_i$ satisfy

$$R_i : X \times A^{\mathcal{N}} \times \Theta \rightarrow \overline{\mathbb{R}}.$$

The input menus $X_i(t_i) \subseteq X_i$ satisfy $X_i(t_i) \supseteq X_i(\hat{t}_i)$ whenever $t_i \succeq \hat{t}_i$. The mechanism induces the following game:

1. nature draws (θ, t) under Γ , and each agent i observes only t_i ;
2. the agents simultaneously send inputs $x_i \in X_i(t_i)$;
3. the mechanism draws $y \sim \rho(x)$ and privately sends y_i to agent i ;

²⁴It is without loss to impose this over the type space $\bigcup \mathbb{T}_i$.

4. the information structure draws $s \sim \pi(\cdot \mid \theta, t)$, and each agent i observes only s_i ;
5. the agents simultaneously choose feasible actions $a'_i \in A_i[t_i]$. Agent i receives $u_i[t_i](\mathbf{a}', \theta) + R_i(x, \mathbf{a}', \theta)$, and the human receives $v(\mathbf{a}', \theta)$.

A pure strategy consists of an input $x_i(t_i)$ and an action rule

$$\alpha_i(t_i, x_i, y_i, s_i) \in A_i[t_i].$$

An equilibrium requires the input and action rule to be optimal for each player $i \in \mathcal{N}$ under the distribution $\beta_i[t_i](d\theta, dt_{-i})\pi(ds \mid \theta, t_i, t_{-i})$.

For a multi-agent indirect mechanism ϕ , let

$$\mathcal{O}(\phi) \subseteq \Delta(\Theta \times \mathbb{T} \times A^{\mathcal{N}})$$

be the set of joint laws of the state, types, and played action profile induced by equilibria of ϕ under $\Gamma(d\theta, dt)\pi(ds \mid \theta, t)$. Indirect mechanisms induce an incomplete information game among players where their ‘reward gradient’ might be shaped by their report. As before, we will work with direct mechanisms.

Definition 4 (Multi-agent direct mechanism). A direct mechanism is a pair (γ, r) , where $r = (r_i)_i$. Its recommendation rule

$$\gamma : \mathbb{T} \rightarrow \Delta\left(\prod_i A^{S_i}\right)$$

assigns each report profile a lottery over profiles of signal-contingent plans. Agent i privately receives its plan $a_i : S_i \rightarrow A$. The reward schedules are

$$r_i : \mathbb{T} \times A^{\mathcal{N}} \times \Theta \rightarrow \overline{\mathbb{R}}.$$

The direct mechanism induces the following game:

1. nature draws (θ, t) under Γ , and each agent i observes only t_i ;
2. the agents simultaneously report $\hat{t}_i \in C_i(t_i)$; write $\hat{t} := (\hat{t}_i)_i$;
3. the mechanism draws a plan profile $\mathbf{a}(\cdot) \sim \gamma(\hat{t})$ and privately sends $a_i(\cdot)$ to agent i ;
4. the information structure draws $s \sim \pi(\cdot \mid \theta, t)$ using the true type profile, and each agent i observes only s_i ;

5. the agents simultaneously choose feasible actions $a'_i \in A_i[t_i]$. Agent i receives $u_i[t_i](\mathbf{a}', \theta) + r_i(\hat{t}, \mathbf{a}', \theta)$, and the human receives $v(\mathbf{a}', \theta)$.

For a plan profile $\mathbf{a}(\cdot)$, write

$$\mathbf{a}(s) := (a_1(s_1), \dots, a_n(s_n))$$

for the induced action profile.

Direct mechanisms simply tell agents what plans—maps from signals to actions—to follow based on their report. The following condition ensures that it is actually optimal to report types honestly and follow the recommended plan.

Definition 5 (Honesty, obedience, and action feasibility). The recommendation rule is *action feasible* if

$$\text{supp marg}_i \gamma(\hat{t}) \subseteq A_i[\hat{t}_i]^{S_i}$$

for every agent i and report profile $\hat{t}$. The direct mechanism is *incentive compatible* if it is action feasible and truthful reporting followed by $a_i(s_i)$ is optimal. Thus, fix an agent i , a type $t_i \in \mathbb{T}_i$, a feasible report $\hat{t}_i \in C_i(t_i)$, and a feasible deviation rule

$$d_i : A^{S_i} \times S_i \rightarrow A_i[t_i].$$

For any plan and signal profiles, let

$$\mathbf{a}^{d_i}(s) := (d_i(a_i(\cdot), s_i), \mathbf{a}_{-i}(s_{-i}))$$

be the action profile after agent i deviates. Incentive compatibility requires

$$\begin{aligned} & \mathbb{E}_{\substack{(\theta, t_{-i}) \sim \beta_i[t_i] \\ s \sim \pi(\cdot | \theta, t_i, t_{-i}) \\ \mathbf{a}(\cdot) \sim \gamma(t_i, t_{-i})}} [u_i[t_i](\mathbf{a}(s), \theta) + r_i((t_i, t_{-i}), \mathbf{a}(s), \theta)] \\ & \geq \mathbb{E}_{\substack{(\theta, t_{-i}) \sim \beta_i[t_i] \\ s \sim \pi(\cdot | \theta, t_i, t_{-i}) \\ \mathbf{a}(\cdot) \sim \gamma(\hat{t}_i, t_{-i})}} [u_i[t_i](\mathbf{a}^{d_i}(s), \theta) + r_i((\hat{t}_i, t_{-i}), \mathbf{a}^{d_i}(s), \theta)]. \end{aligned}$$

Honest and obedient play induces the outcome

$$O(\gamma, r) \in \Delta \left(\Theta \times \mathbb{T} \times A^{\mathcal{N}} \right),$$

under $(\theta, t) \sim \Gamma$, $s \sim \pi(\cdot | \theta, t)$, and $\mathbf{a}(\cdot) \sim \gamma(t)$, with played action profile $\mathbf{a}(s)$. Let

$$\mathcal{O}_n := \bigcup_{\phi} \mathcal{O}(\phi),$$

and let $\mathcal{O}_n^{\text{IC}}$ be the set of outcomes $O(\gamma, r)$ induced by incentive-compatible direct mechanisms.

Proposition 6 (Multi-agent revelation principle). $\mathcal{O}_n = \mathcal{O}_n^{\text{IC}}$.

Proof of Proposition 6. We extend the single-agent construction and preserve correlation across private mechanism outputs. Fix an indirect mechanism and a pure-strategy equilibrium (x, α) , and write $x(\hat{t}) := (x_i(\hat{t}_i))_i$. For each type report $\hat{t}_i$ and mechanism output y_i , define the plan

$$a_{i, \hat{t}_i, y_i}(s_i) := \alpha_i(\hat{t}_i, x_i(\hat{t}_i), y_i, s_i).$$

For report profile $\hat{t}$, let $\gamma(\hat{t})$ be the joint law of these plans under $y \sim \rho(x(\hat{t}))$, and set

$$r_i(\hat{t}, \mathbf{a}', \theta) := R_i(x(\hat{t}), \mathbf{a}', \theta).$$

Equilibrium feasibility gives action feasibility, and honest, obedient play reproduces the indirect outcome.

A direct deviation depends on agent i 's plan and realized physical signal. The indirect agent can copy it after sending $x_i(\hat{t}_i)$, which is available because $\hat{t}_i \in C_i(t_i)$ and the input menus are monotone. The mechanism output y_i determines the same plan, and the agent later observes the same s_i under the true physical kernel $\pi(\cdot \mid \theta, t)$. The indirect equilibrium therefore blocks every direct deviation. This proves $\mathcal{O}_n \subseteq \mathcal{O}_n^{\text{IC}}$.

For the reverse inclusion, use $X_i = \mathbb{T}_i$ with input menus $X_i(t_i) = C_i(t_i)$, $Y_i = A^{S_i}$, $\rho = \gamma$, and $R_i = r_i$. Direct incentive compatibility makes truthful reporting and obedience an equilibrium. $\square$

The logic of Proposition 6 is straightforward and follows the usual arguments for the revelation principle. The non-standard aspects are (i) the set of types that can be misreported as another type is constrained by the verification order a la [Green and Laffont \(1986\)](#); and (ii) recommendations are plans that map ex-interim signal realizations to actions. Thus our designer is, in the language of [Bergemann and Morris \(2019\)](#), non-omniscient and, what is more, the AI learns more about the state ex-interim. We think these features capture how humans might interact with AI agents in practice.

5 APPLICATIONS: MULTI-AGENT

We study three ways in which one AI can help discipline another. Co-player scoring uses correlation across reports when ground truth is unavailable. Coupled reward design uses the difference between agents' actions to reveal linked private biases.

Weak-to-strong oversight instead lets a task-ignorant monitor use a private diagnosis of the acting AI’s bias to choose among increasingly rich control instruments: binary approval, a permission set, or an action reward. These mechanisms are counterparts of the *scalable oversight* agenda in AI safety (Christiano et al., 2018; Leike et al., 2018; Bowman et al., 2022); Hammond et al. (2025) survey the risks specific to multi-agent deployments.

Example III (Peer discipline). When agents’ private information is correlated, one agent’s type is informative about its co-players’ types. We show how mechanisms can use this information to discipline actions. We will first establish this in a general environment, then specialize it to two types and two states to illustrate the basic forces.

Arbitrary types. Assume that the marginal over types under Γ is finitely supported. For each agent i , let

$$T_i^0 := \text{suppmarg}_{\mathbb{T}_i} \Gamma \quad \text{and} \quad T_{-i}^0 := \prod_{j \neq i} T_j^0.$$

Write $\beta_i[t_i](t_{-i})$ for type t_i ’s marginal probability of the co-player type profile t_{-i} . We assume that for each player i , each $t_i \in T_i^0$ assigns probability 1 to T_{-i}^0 .²⁵ We will also set $C_i(t_i) = \mathbb{T}_i$ i.e., every report is feasible.

Implementation requires honesty and obedience only for types $t_i \in T_i^0$ i.e., those that are in the support of the principal’s belief Γ . It will also be without loss to restrict our attention to reports in T_i^0 : the mechanism can simply ‘assign reward $-\infty$ ’ after every report that lies outside the support of Γ .

Now consider the separation condition

$$\beta_i[t_i] \neq \beta_i[\hat{t}_i] \tag{S}$$

for every agent i and every pair of distinct supported reports $t_i, \hat{t}_i \in T_i^0$.

For an action-feasible rule

$$\gamma : \mathbb{T} \rightarrow \Delta \left(\prod_i A^{S_i} \right),$$

we let $\gamma_\theta(t)$ be the distribution of the played action profile under $\gamma(t)$ and the information structure π . For some $\Lambda > 0$, choose the following reward schedule:

$$r_i(\hat{t}, \mathbf{a}', \theta) = \begin{cases} \Lambda \left[2\beta_i[\hat{t}_i](\hat{t}_{-i}) - \|\beta_i[\hat{t}_i]\|_2^2 \right] - u_i[\hat{t}_i](\mathbf{a}', \theta), & \begin{matrix} \hat{t}_j \in T_j^0 \text{ for every } j, \\ \mathbf{a}' \in \text{supp} \gamma_\theta(\hat{t}), \end{matrix} \\ -\infty, & \text{otherwise.} \end{cases} \tag{PR}$$

²⁵This is a standard ‘belief closure’ condition which is implied by (but much weaker than) a common prior. It simplifies the exposition but is not necessary though the mechanism becomes more complicated.

where (PR) stands for peer reward. The first term scores the report against co-player reports. The second cancels the reported utility on the support of actions the designer wishes to implement.

Proposition 7 (Support-wise implementability of every feasible rule). Fix any action-feasible rule γ . If $n \geq 2$ and (S) holds, then, for sufficiently large $\Lambda > 0$, the schedule (PR) implements γ on the support of Γ .

Proof of Proposition 7. For a type $t_i \in T_i^0$ in the support of Γ , truthful reporting and obedience give a finite payoff: the utility terms cancel and the quadratic score is finite. Every action outside the target support gives $-\infty$, so obedience is optimal. Every report outside the support of Γ also gives $-\infty$ after every action and is therefore strictly worse.

It remains to compare misreports that are supported on Γ . If type t_i instead reports a distinct $\hat{t}_i \in T_i^0$, the quadratic-score identity in Equation (PR) gives the expected score loss

$$\Lambda \|\beta_i[t_i] - \beta_i[\hat{t}_i]\|_2^2.$$

If no agent i has two supported reports we are done. Otherwise, finiteness and (S) imply

$$\delta := \min_{\substack{i, t_i, \hat{t}_i \in T_i^0 \\ \hat{t}_i \neq t_i}} \|\beta_i[t_i] - \beta_i[\hat{t}_i]\|_2^2 > 0.$$

Because A , Θ , and the supported report sets are finite and utilities are real-valued,

$$M := \max_{\substack{i, t_i, \hat{t}_i \in T_i^0, \hat{t}_i \neq t_i \\ \hat{t}_{-i} \in T_{-i}^0, \theta \in \Theta \\ \mathbf{a} \in \text{supp } \gamma_\theta(\hat{t})}} |u_i[t_i](\mathbf{a}, \theta) - u_i[\hat{t}_i](\mathbf{a}, \theta)| < \infty.$$

Any supported false report that yields a finite payoff lowers the expected score by at least $\Lambda\delta$ and raises the remaining utility term by at most M . Its payoff gain is therefore at most $M - \Lambda\delta$. Any $\Lambda \geq M/\delta$ makes truth-telling optimal for every supported type. Honest and obedient play then induces γ under Γ . $\square$

A binary state. We now illustrate Proposition 7 with a simple two agents and two state example. $\mathcal{N} = \{1, 2\}$ and $\theta \in \{0, 1\}$ with $\mathbb{P}(\theta = 1) = 1/2$. If $\theta = 1$, both agents have type H . If $\theta = 0$, the type profile is (H, L) or (L, H) with equal probability. The types are the agents' private signals i.e., there is no further interim learning. Their joint distribution is

	$t_2 = H$	$t_2 = L$
$t_1 = H$	1/2	1/4
$t_1 = L$	1/4	0.

Each agent chooses $a_i \in \{0, 1\}$. The human wants the agents to agree on the state, while each agent prefers the high action. Hence, payoffs are

$$v(\mathbf{a}, \theta) = \mathbb{1}\{a_1 = a_2 = \theta\} \quad \text{and} \quad u_i(\mathbf{a}, \theta) = a_i.$$

The human therefore wants both agents to take action 1 if and only if $t = (H, H)$ and this gives the human expected payoff equal to one. Write $a^*(\hat{t}) = 1$ if $\hat{t} = (H, H)$ and $a^*(\hat{t}) = 0$ otherwise.

Now observe that types H and L have the following belief about their co-players' type:

$$(\beta_iH, \beta_i[H](L)) = \left(\frac{2}{3}, \frac{1}{3}\right) \quad \text{and} \quad (\beta_i[L](H), \beta_iL) = (1, 0).$$

The unscaled quadratic score in (PR) is therefore

$q(\hat{t}_i, \hat{t}_j)$	$\hat{t}_j = H$	$\hat{t}_j = L$
$\hat{t}_i = H$	7/9	1/9
$\hat{t}_i = L$	1	-1.

For $j \neq i$ and any $\Lambda > 0$, set

$$r_i(\hat{t}, \mathbf{a}, \theta) = \begin{cases} \Lambda q(\hat{t}_i, \hat{t}_j) - a_i, & \mathbf{a} = (a^*(\hat{t}), a^*(\hat{t})), \\ -\infty, & \text{otherwise.} \end{cases}$$

The term $-a_i$ cancels the agent's preference for action 1, and the off-target penalty enforces the common action. An H -type obtains expected score 5/9 from reporting H and 1/3 from reporting L . An L -type obtains score 1 from reporting L and 7/9 from reporting H . Truth raises either type's score by 2/9, so the schedule implements the human's favorite rule to get a maximum payoff of 1.

Because we work on the universal type space, the first-order beliefs $\beta_i[t_i]$ over co-player reports might contain higher-order beliefs in the spirit of (Miller et al., 2005; Prelec, 2004). This procedure is in the spirit of work by Qiu et al. (2026) who show both theoretically and empirically that using scoring rules can elicit truthful reporting among language models under a common prior. They show that peer-score-ranked answer pairs for direct preference optimization can recover most of the accuracy lost to deceptive fine-tuning. Importantly, they also find that peer scoring performs better when the participant models are larger relative to the expert model.

Our broad conceptual point is quite similar, though it is worth highlighting some contrasts. First, our designer (by the construction of types) knows type t_i 's beliefs $\beta_i[t_i] \in \Delta(\Theta \times \mathbb{T}_{-i})$ and our procedure allows for non-common priors.²⁶ By contrast,

²⁶We also focus on disciplining actions (what the designer values) rather than eliciting truthful information though this is simple in the presence of unbounded rewards.

Qiu et al. (2026) do not give the designer this conditional-belief map. They instead use an expert to supply $\mathbb{P}(A_j \mid A_i)$ —the expert predicted probability of player j ’s answer given player i ’s answer—that is, in turn, used to score i ’s answer. Second, our procedure also uses peer prediction to elicit agents’ capabilities and preferences—not just information about the fundamental state $\theta \in \Theta$. ♦

Example IV (Competition through coupled rewards). We now show how competition can make every action close to the human’s ideal action, even when the human does not know the agents’ biases.

Setup. A human faces two agents. Nature draws $(b_1, b_2) \sim \Gamma$, and agent i observes its own bias b_i . The human observes neither bias. It is common knowledge that the biases lie on a curve:

$$\text{supp}\Gamma \subseteq \mathcal{C}_f := \{(b_1, b_2) \in \mathbb{R}^2 : b_2 = f(b_1)\}, \quad (\text{Curve})$$

where $f : \mathbb{R} \rightarrow \mathbb{R}$ is continuous and strictly decreasing, with $f(0) = 0$. Thus the agents have countervailing biases away from the origin. Because f is one-to-one, each agent can infer the other agent’s bias from its own. Assume also that $\mathbb{E}_\Gamma[(b_2 - b_1)^2] < \infty$.

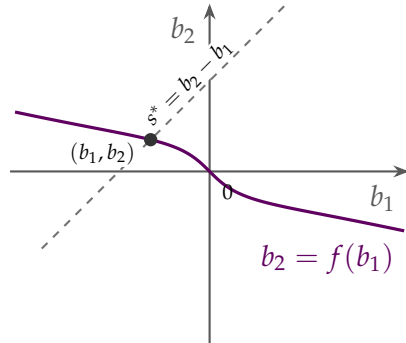
Both agents observe the state θ and simultaneously choose actions $a_1, a_2 \in \mathbb{R}$. The human and agent i have payoffs

$$v(\mathbf{a}, \theta) = -\sum_{i=1}^2 (a_i - \theta)^2 \quad \text{and} \quad u_i(a_i, \theta) = -(a_i - \theta - b_i)^2.$$

Thus every action enters the human’s payoff, while each agent cares about its own action.

Figure 9 shows the nonlinear bias curve and the best-responses without the mechanism. Since agent i ’s payoff is independent of j ’s action, the best response curve is flat/vertical.

(a) The action difference reveals bias



(b) Without coupled rewards

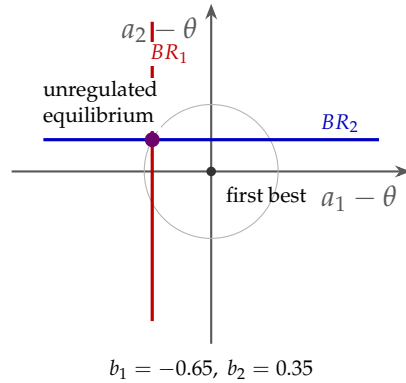


Figure 9: Biases without coupled rewards

Coupled rewards. We now introduce dependence in the reward schedule such that player i 's incentives are shaped by j 's actions. Write $D := a_2 - a_1$ for the difference between the actions. The map $h(x) := f(x) - x$ is continuous, strictly decreasing, and onto $\mathbb{R}$; let $g := h^{-1}$ denote its inverse. For any $\eta > 0$, give the agents the rewards

$$\begin{aligned} r_1(D) &= 2\eta \int_0^{D/\eta} g(s) \, ds + \frac{D^2}{2}, \\ r_2(D) &= -2\eta \int_0^{D/\eta} (g(s) + s) \, ds + \frac{D^2}{2}. \end{aligned} \tag{CR}$$

These schedules depend only on the known curve. The parameter η controls the size of the action difference that reveals the realized bias pair. Agent i receives $u_i(a_i, \theta) + r_i(D)$.

Proposition 8 (Coupled rewards make loss arbitrarily small). If condition (Curve) holds, then, for every $\eta > 0$, the rewards in (CR) induce a unique Nash equilibrium for every realized bias pair and state. The equilibrium actions are

$$a_1^* = \theta - \frac{\eta}{2}(b_2 - b_1) \quad \text{and} \quad a_2^* = \theta + \frac{\eta}{2}(b_2 - b_1).$$

The loss is $L^* = \frac{\eta^2}{2}(b_2 - b_1)^2$ hence $\mathbb{E}_\Gamma[L^*] \rightarrow 0$ as $\eta \downarrow 0$.

Proof. We first verify concavity. We then show how the difference between the actions reveals the two biases. Since g is strictly decreasing, the first integral in (CR) is concave in D . Since $g(s) + s = f(g(s))$ is strictly increasing, the negative integral term in r_2 is concave in D . Holding the other action fixed, the intrinsic quadratic has second derivative -2 , while $D^2/2$ has second derivative 1. Each agent's payoff is therefore strictly concave in its own action, so its first-order condition is necessary and sufficient.

Let $s := D/\eta$. Differentiating (CR) gives

$$r_1'(D) = 2g(s) + D, \quad r_2'(D) = -2(g(s) + s) + D.$$

The two first-order conditions reduce to

$$a_1 - \theta = b_1 - g(s) - \frac{D}{2}, \quad a_2 - \theta = b_2 - g(s) - s + \frac{D}{2}.$$

Because $D = (a_2 - \theta) - (a_1 - \theta)$, subtracting these equations gives $s = b_2 - b_1$. Condition (Curve) then gives

$$s = f(b_1) - b_1 = h(b_1), \quad g(s) = b_1.$$

It follows that

$$D^* = \eta(b_2 - b_1), \quad a_1^* - \theta = -\frac{D^*}{2}, \quad a_2^* - \theta = \frac{D^*}{2}.$$

The first-order conditions have a unique solution, and strict concavity makes it the unique Nash equilibrium. Substitution gives the stated loss L^* and we can take expectation and send η to zero. $\square$

Coupling induces competition. Coupling rewards to penalize disagreement pushes the agents toward similar actions. But this, *per se*, need not push them toward choosing $a_i = \theta$; they might simply choose the same fixed action. The schedule (CR) works because the schedule simultaneously (i) reveals agents' biases; and (ii) correct agents' marginal incentives. To see this plainly, observe agents' first-order conditions are

$$a_1 - \theta = b_1 - g(D/\eta) - \frac{D}{2} \quad \text{and} \quad a_2 - \theta = b_2 - g(D/\eta) - \frac{D}{\eta} + \frac{D}{2}.$$

Subtracting these conditions and using $D = a_2 - a_1$ gives

$$\frac{D}{\eta} = b_2 - b_1$$

in equilibrium. Because $b_2 = f(b_1)$ and $g = [f(x) - x]^{-1}$, it follows that

$$g(D/\eta) = b_1 \quad \text{and} \quad g(D/\eta) + D/\eta = b_2.$$

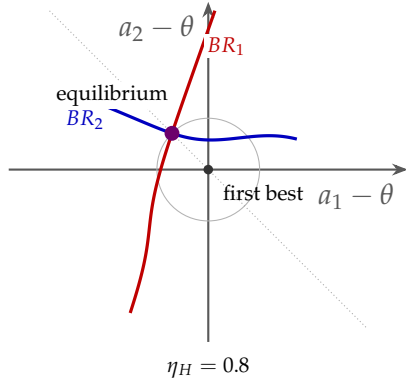
Substituting these identities back into the first-order conditions cancels the realized biases and gives

$$a_1 - \theta = -\frac{D}{2} \quad \text{and} \quad a_2 - \theta = \frac{D}{2}.$$

Thus the reward schedule uses the equilibrium disagreement to infer each bias and splits the remaining action errors symmetrically around θ .

Figure 10 plots these best responses for two values of η . Each red curve contains the action pairs that satisfy agent 1's first-order condition. Each blue curve does the same for agent 2. The two curves cross once, at the equilibrium in **Proposition 8**. As η is smaller, it preserves the relation $D^*/\eta = b_2 - b_1$ but moves both actions towards θ .

(a) Higher η : more disagreement



(b) Lower η : less disagreement

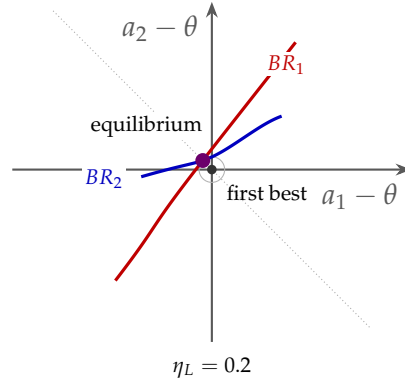


Figure 10: Coupled rewards to steer best responses

We emphasize that although the reward schedule can be thought of as eliciting the bias before correcting it, this procedure is fully simultaneous. The reason is that a schedule (CR) with any $\eta > 0$ induces an equilibrium outcome that reveals the bias difference since $(a_2^* - a_1^*)/\eta = b_2 - b_1$. Moreover, because of our assumption that biases lie on a lower-dimensional manifold $\mathcal{C}_f$, this can exactly identify b_1 and b_2 which can then enter agents' marginal rewards to correct their incentives. Our construction is in the spirit of [Koh and Martinez-Bruera \(2025\)](#) who exploit a homeomorphism between a random network of externalities and the corresponding efficient allocation to construct nonlinear taxes that can implement the first best outcome realization-by-realization. Here, biases are supported on the curve f which makes the action difference reveal the bias pair. But that difference is costly because both actions enter human welfare. Our mechanism thus uses the curvature of reward (taking η small) to approach the first best. Of course, biases might, in practice, be supported on the two-dimensional plane so the reward schedule can no longer be designed to achieve the first-best outcome. We think an interesting open problem is how reward should be optimally coupled then—and more broadly whether such principles might scale to more complicated environments.

Our result is related in spirit to the communication game of [Collina et al. \(2025\)](#) who show how competition among biased agents can induce truth-telling in a persuasion game (see also [Battaglini \(2002\)](#) for an early important result). Here, the point is to shape agents' actions and the point is how rewards can be coupled to discipline behavior. ♦

Example V (Scalable oversight and reward shaping). Finally, we show how our framework can shed light on the increasingly common practice of using a weak but trusted or interpretable monitor to control or reward a strong but untrusted or hard-to-interpret actor.

Setup. The human faces a strong actor S and a weak monitor W . The state is

$\theta \sim U[0, 1]$, and the strong actor's bias is the random variable $b \sim \tau$, where $\mathbb{E}[b^2] < \infty$. The state and bias of the strong actor are independent. The monitor W has a known bias $w \in \mathbb{R}$.

Before the state is realized, the weak monitor W privately observes the bias b of the strong actor S . This might be because the weak monitor, despite its inability to do what the strong actor does, might be able to monitor the chain of thought and infer misaligned behavior. The strong actor knows its bias and later observes the state θ perfectly in the interim stage. The implemented action lies in $\mathbb{R}$. The players' payoffs are:

- **Human:** $-(a - \theta)^2$.
- **Strong actor S :** $-(a - \theta - b)^2$.
- **Weak monitor W :** $-(a - \theta - w)^2$.

Before the bias is realized, the human chooses one of three oversight regimes. After privately observing b , the weak monitor chooses an instrument from that regime. The strong actor then observes the instrument and the state θ and acts.

Some notation is in order to formalize the set of instruments the monitor W can choose. Call a state-independent reward $r : \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty\}$ *admissible* if

$$\operatorname{argmax}_{a \in \mathbb{R}} \left\{ -(a - \theta - b)^2 + r(a) \right\}$$

is nonempty for every $(\theta, b) \in [0, 1] \times \operatorname{supp}(\tau)$.²⁷ Let $\mathcal{R}$ denote the set of admissible rewards. A special case of rewards is when actions are either allowed or not. A nonempty closed set $D \subseteq \mathbb{R}$ can be thought of as a permission set that simply corresponds to

$$r(a) = \begin{cases} 0 & \text{if } a \in D \\ -\infty & \text{otherwise.} \end{cases}$$

Under reward design, the weak monitor's strategy maps the strong actor's bias b into a reward schedule in $\mathcal{R}$.

- *Regime I: binary approval.* The weak monitor chooses between approval and rejection. Approval gives the strong actor the permission set $D = \mathbb{R}$. Rejection gives it the singleton permission set $D = \{1/2\}$ and therefore implements the human's optimal fixed action.

²⁷We use an arbitrary but fixed tiebreaking rule. The instruments constructed below induce a unique action almost surely.

- *Regime II: permission-set delegation.* The weak monitor chooses any nonempty closed permission set $D \subseteq \mathbb{R}$. The strong actor then chooses its preferred action in D .
- *Regime III: reward design.* The weak monitor chooses any admissible state-independent reward r . The strong actor then maximizes $-(a - \theta - b)^2 + r(a)$ over $a \in \mathbb{R}$. Rewards do not enter human welfare.

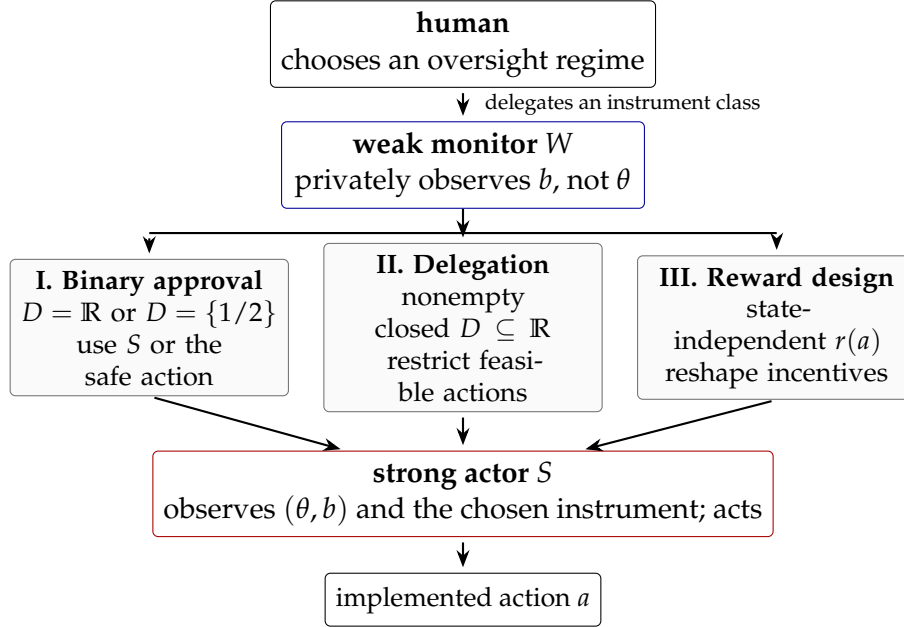


Figure 11: Illustration of oversight game

Figure 11 illustrates the game. Notice that regimes are nested. Regime I restricts the monitor to two permission sets, so it is a special case of Regime II. Since delegation is a special case of rewards in which the image is $\{0, -\infty\}$, Regime II is therefore a special case of Regime III.

Regime I: binary approval. Approval makes S choose $a = \theta + b$ and gives W loss $(b - w)^2$. Rejection gives W loss $1/12 + w^2$. Thus, breaking its ties in favor of approval, the monitor approves if and only if

$$(b - w)^2 \leq \frac{1}{12} + w^2.$$

The monitor uses its diagnosis only to decide whether to use the strong actor S but, if it does, cannot shape how S acts. This is broadly analogous to trusted-monitor control, where a monitor decides whether to use an untrusted model's output or defer to a trusted fallback (Greenblatt et al., 2024c; Xiao and Phuong, 2026).

Regime II: delegation. The monitor now controls the feasible actions. One optimal representative is

$$D^{II}(b) = \begin{cases} (-\infty, 1 + 2w - b], & 0 < b - w < 1/2, \\ \mathbb{R}, & b = w, \\ [2w - b, \infty), & -1/2 < b - w < 0, \\ \{w + 1/2\}, & |b - w| \geq 1/2. \end{cases}$$

When S wants a larger action than W , the monitor imposes a cap. When S wants a smaller action, it imposes a floor. Once relative bias reaches $1/2$, the monitor removes all state discretion and implements its own optimal fixed action $w + 1/2$. Fine-grained privilege-control systems use the same logic when they restrict permissible tool calls and specify fallbacks for blocked calls (Shi et al., 2025).

Regime III: reward design. The affine schedule

$$r_b^{III}(a) = -2(b - w)a$$

makes the strong actor choose $a = \theta + w$. This action is first best for W , so no other reward can improve its payoff. This regime is in the spirit of scalable reward modeling: a weaker system helps shape the reward signal that trains or steers a stronger system when direct human evaluation is hard (Leike et al., 2018; Baker et al., 2025). Variations of these ideas have been recently used. For instance, Kenton et al. (2026) use second model critiques the acting model before an LLM judge assigns the training reward and find that this design reduces reward hacking. Our regime is in this spirit, and asks how AI-provided feedback changes the incentives of a stronger model.

Proposition 9 (Losses across oversight regimes). Conditional on the realized bias b , the monitor's losses are

$$L_W^I(b) = \min \left\{ (b - w)^2, \frac{1}{12} + w^2 \right\},$$

$$L_W^{II}(b) = \begin{cases} |b - w|^2 - \frac{4}{3}|b - w|^3, & |b - w| < 1/2, \\ \frac{1}{12}, & |b - w| \geq 1/2, \end{cases}$$

$$L_W^{III}(b) = 0.$$

The human's losses are

$$L_H^I(b) = \begin{cases} b^2, & (b - w)^2 \leq \frac{1}{12} + w^2, \\ \frac{1}{12}, & (b - w)^2 > \frac{1}{12} + w^2, \end{cases}$$

$$L_H^{II}(b) = L_W^{II}(b) + 2w(b - w)(1 - 2|b - w|)_+ + w^2,$$

$$L_H^{III}(b) = w^2,$$

where $x_+ := \max\{x, 0\}$. Ex-ante losses are the expectations of these expressions over $b \sim \tau$.

Proof of Proposition 9. For this proof, write $\delta := b - w$. We solve each regime. In Regime I, approval yields the monitor error δ , while rejection yields the error $1/2 - \theta - w$. Their squared expectations are δ^2 and $1/12 + w^2$. The human losses from the same choices are b^2 and $1/12$. This proves the Regime-I formulas under the stated tie-break.

For Regime II, translate actions by the monitor's bias: set $y = a - w$. The monitor wants $y = \theta$, while the strong actor wants $y = \theta + \delta$. The one-type interval reduction used in the proof of Proposition 4 applies pointwise. If $\delta > 0$, extending an interval's lower end cannot hurt the monitor, so we can optimize an upper cap $y \leq c$. Its loss has derivative

$$\frac{dL(c)}{dc} = \begin{cases} 2c - 1, & c \leq \delta, \\ \delta^2 - (1 - c)^2, & \delta < c < 1 + \delta, \\ 0, & c \geq 1 + \delta. \end{cases}$$

Thus $c = 1 - \delta$ is optimal when $0 < \delta < 1/2$, while $c = 1/2$ is optimal when $\delta \geq 1/2$. Reflecting both the state and action gives the floor for $\delta < 0$. Translating back by w

gives the displayed permission sets. Direct integration gives

$$L_W^{II}(b) = \begin{cases} |\delta|^2 - \frac{4}{3}|\delta|^3, & |\delta| < 1/2, \\ \frac{1}{12}, & |\delta| \geq 1/2, \end{cases} \quad \mathbb{E}[a - \theta - w \mid b] = \delta(1 - 2|\delta|)_+.$$

Since $a - \theta = (a - \theta - w) + w$, expanding the square gives

$$L_H^{II}(b) = L_W^{II}(b) + 2w\delta(1 - 2|\delta|)_+ + w^2.$$

In Regime III, strict concavity under r_b^{III} gives the unique action $a = \theta + w$. The monitor obtains zero loss and the human obtains loss w^2 . $\square$

Proposition 9 compares payoffs for the human and weak monitor across three regimes. Since the instrument classes are nested and the weak monitor's payoff depends only on the action of the strong actor S and the state, we have

$$L_W^I \geq L_W^{II} \geq L_W^{III} = 0.$$

But this ordering need not hold for the human—giving the weak monitor W more freedom to shape the rewards or permissions of the strong actor S need not improve payoffs precisely because W might, itself, be biased.²⁸ The basic tradeoff here is familiar to the overarching theme of delegation: giving W more discretion allows it to use its superior information about the strong agent's bias more precisely, and tailor rewards and permissions. But W shapes the behavior of S to its *own* preference which, itself, is biased. Thus, the optimal regime must depend on the relative strengths of these biases.

²⁸Similar concerns were raised by the Roman author Juvenal; see [Hurwicz \(2008\)](#).

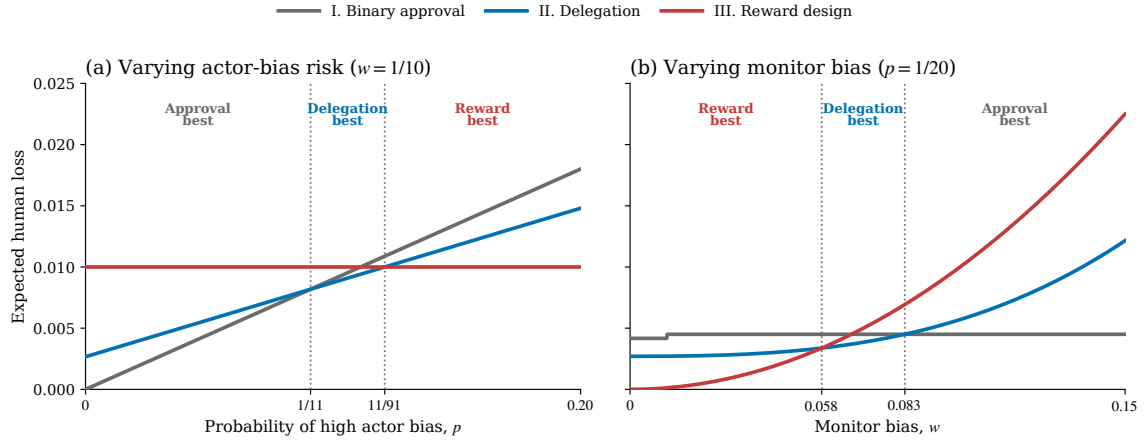


Figure 12: **Each authority regime can be optimal for the human.** In both panels, the strong actor has bias $3/10$ with probability p and zero bias otherwise. Panel (a) fixes $w = 1/10$ and varies p . Panel (b) fixes $p = 1/20$ and varies w . The dotted lines mark the values at which the human’s preferred regime changes.

We illustrate with the simplest numerical example. Let the strong actor’s bias take values $b \in \{0, 3/10\}$ and let $p = \mathbb{P}(b = 3/10)$. Let the weak monitor’s known bias be $w > 0$. The human’s expected losses are depicted in Figure 12. Panel (a) fixes the weak monitor’s bias at $w = 1/10$ and varies the probability that the strong actor is biased. We see that the lower-envelope is attained by different regimes, depending on how biased the strong actor is. As the probability that the strong actor is biased increases, we want to give the monitor more discretion to shape its rewards. Panel (b) fixes the distribution over the strong actor’s bias and varies the monitor’s (known) bias. Here the opposite logic holds—as the weak monitor becomes more biased, we want to give it less ability to shape rewards.

Optimal oversight. Regimes I–III are specific oversight mechanisms. More generally, an oversight mechanism can be thought of as a set of reward schedules $\mathcal{M} \subseteq \mathcal{R}$ from which the weak monitor chooses.²⁹ Recall that $\mathcal{R}$ contains the reward schedules $r : \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty\}$ under which the strong actor has an optimal action. For instance, Regime III allowed the monitor to pick any schedule in $\mathcal{R}$ and always implemented $a = \theta + w$ which is the weak monitor’s first best.

We now construct a family $\mathcal{M}^* \subseteq \mathcal{R}$ that attains the human first best. Our construction will not require the human to know the monitor’s bias w ; it only uses a bound $\bar{w} \geq |w|$. Fix a pair of extreme actions $\bar{a}$ and $1 - \bar{a}$ with $\bar{a} > 1 + \bar{w} + \sqrt{\bar{w}^2 + 2\bar{w}}$. We

²⁹Thus, $\mathcal{R}$ can be thought of as W ’s action space, and in our framework the (human) mechanism designer delegates a subset.

construct $\mathcal{M}^* = \{r_{\hat{b}} : \hat{b} \in \text{supp}(\tau)\}$, where

$$r_{\hat{b}}(a) = \begin{cases} -2\hat{b}a, & a \in [0, 1], \\ (\bar{a} - 1)^2 - 2\bar{a}\hat{b}, & a = \bar{a}, \\ (\bar{a} - 1)^2 - 2(1 - \bar{a})\hat{b}, & a = 1 - \bar{a}, \\ -\infty, & \text{otherwise.} \end{cases}$$

Proposition 10 (Robust first best). Suppose the monitor's bias w has a known bound $|w| \leq \bar{w}$. Under $\mathcal{M}^* \subseteq \mathcal{R}$, for every $w \in [-\bar{w}, \bar{w}]$ and every realization of the strong actor's bias b , the weak monitor uniquely chooses a reward $r \in \mathcal{M}^*$ that induces $a = \theta$ almost surely.

Proposition 10 states that whenever the weak monitor's bias is known to lie within some bounds, we can pick the menu of reward schedules in such a way that implements the first-best action $a = \theta$. Three reward schedules $r_{\hat{b}} \in \mathcal{M}^*$ are illustrated in **Figure 13**.

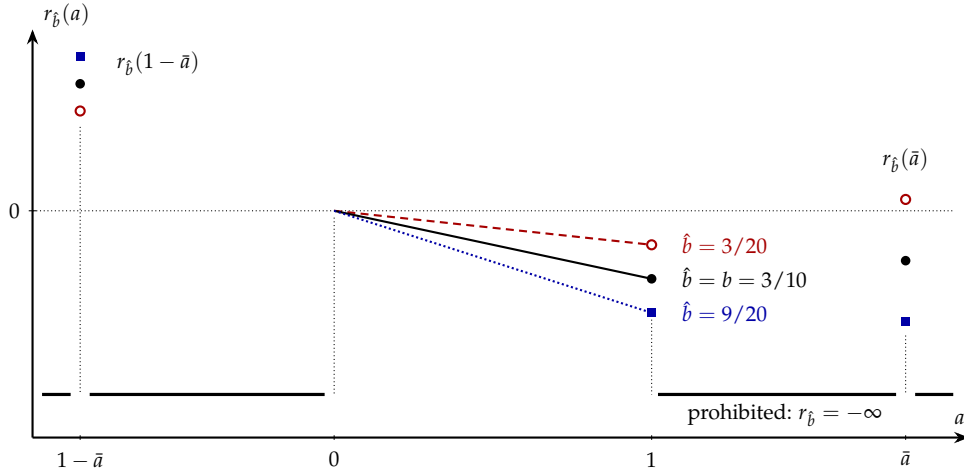


Figure 13: The reward menu $\mathcal{M}^*$ ($\bar{a} = 9/5$; normalized so that $r_{\hat{b}}(0) = 0$)

Each schedule $r_{\hat{b}} \in \mathcal{M}^*$ has three parts with distinct roles. First, the linear tilt of slope $-2\hat{b}$ over $[0, 1]$ cancels a bias of exactly $\hat{b}$ which, recall, is W 's report of S 's bias. To see this, we can complete the square:

$$-(a - \theta - b)^2 + r_{\hat{b}}(a) = -(a - \theta - (b - \hat{b}))^2 + \text{terms independent of } a.$$

Hence, facing the reward schedule $r_{\hat{b}}$, a strong actor with true bias b behaves as if its bias were the *residual* $b - \hat{b}$ i.e., it chooses the truncation of $\theta + (b - \hat{b})$. Hence,

if the weak monitor W reports truthfully, this debiases the strong actor fully and implements $a = \theta$.

Second, the reward at the isolated extreme action $\bar{a}$ is calibrated so that, comparing $\bar{a}$ with the top permitted regular action $a = 1$,

$$\underbrace{\left[-(\bar{a} - \theta - b)^2 + r_{\hat{b}}(\bar{a})\right]}_{S's \text{ payoff from choosing } \bar{a}} - \underbrace{\left[-(1 - \theta - b)^2 + r_{\hat{b}}(1)\right]}_{S's \text{ payoff from choosing } a = 1} = 2(\bar{a} - 1)(\theta + (b - \hat{b}) - 1).$$

The actor therefore takes $\bar{a}$ exactly when its residual ideal point $\theta + (b - \hat{b})$ exceeds 1—which never happens after a truthful report. This happens precisely when the monitor W understates its bias ($\hat{b} < b$) and the state is high ($\theta > 1 - (b - \hat{b})$).

Third, the reward at the other isolated extreme action $1 - \bar{a}$ mirrors this: comparing $1 - \bar{a}$ with the bottom permitted regular action $a = 0$,

$$\underbrace{\left[-(1 - \bar{a} - \theta - b)^2 + r_{\hat{b}}(1 - \bar{a})\right]}_{S's \text{ payoff from choosing } 1 - \bar{a}} - \underbrace{\left[-(\theta + b)^2 + r_{\hat{b}}(0)\right]}_{S's \text{ payoff from choosing } a = 0} = -2(\bar{a} - 1)(\theta + (b - \hat{b})).$$

The actor therefore takes $1 - \bar{a}$ exactly when its residual ideal point $\theta + (b - \hat{b})$ falls below 0. This happens precisely when the monitor W *overstates* its bias ($\hat{b} > b$) and the state is low ($\theta < \hat{b} - b$). **Figure 14** depicts the behavior the schedules induce in the strong agent, and the monitor's payoff from each schedule.

The basic reason the reward schedules include these extreme actions $\bar{a}$ and $1 - \bar{a}$ is to counteract the weak monitor's temptation to under or over-report the strong actor's bias. First suppose we had neither extreme action but the reward schedule retained the linear piece. Then, a weak monitor with bias $w > 0$ would like to understate the strong actor's bias and report $\hat{b} = b - w$. Doing so would leave the strong actor with residual bias w so the weak monitor obtains its first best $a = \theta + w$. This was exactly the schedule chosen by the weak monitor in Regime III. Symmetrically, a monitor with $w < 0$ would like to overstate and report $\hat{b} = b - w > b$. The way to deter these deviations is to force the reward schedule to tempt the strong actor with these distant actions— $\bar{a}$ to deter understatement, and $1 - \bar{a}$ to deter overstatement. This induces a strong actor with bias b to take these extreme actions when facing reward schedule $r_{\hat{b}}$ where $\hat{b} \neq b$.

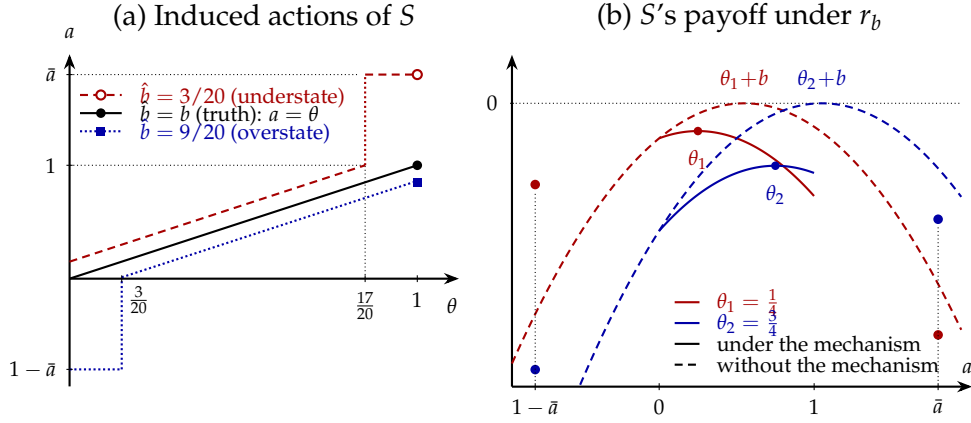


Figure 14: Strong actor's best response and payoffs

Notice that the set $\mathcal{M}^*$ uses no information about the monitor's bias beyond the bound $\bar{w}$. This is because incentives here are asymmetric: the monitor's gain from misreporting is at most of order $|w|$, while its loss from the strong actor taking an extreme action $\in \{\bar{a}, 1 - \bar{a}\}$ is of order $\bar{a}^2$. Hence, sufficiently extreme actions can deter every monitor with $|w| \leq \bar{w}$.

Proof of Proposition 10. Each schedule is admissible because its finite-valued restriction is continuous on the compact set $\{1 - \bar{a}\} \cup [0, 1] \cup \{\bar{a}\}$, so the actor's payoff attains a maximum. Fix a true bias $b \in \text{supp}(\tau)$ and a report $\hat{b} \in \text{supp}(\tau)$. By the three displays above, the actor chooses the truncation of $\theta + (b - \hat{b})$ to $[0, 1]$ when $\theta + (b - \hat{b}) \in [0, 1]$, chooses $\bar{a}$ when $\theta + (b - \hat{b}) > 1$, and chooses $1 - \bar{a}$ when $\theta + (b - \hat{b}) < 0$.

If $\hat{b} = b$, the actor chooses $a = \theta$ at every state except $\theta \in \{0, 1\}$, where it is indifferent with an extreme action; these are null events. The monitor's expected loss is w^2 .

If $\hat{b} < b$, write $\delta := b - \hat{b}$ and suppose first that $\delta \leq 1$. The low extreme action is never taken: the actor chooses $\theta + \delta$ on $\theta < 1 - \delta$ and $\bar{a}$ on $\theta > 1 - \delta$, so the monitor's expected loss in excess of w^2 is

$$\begin{aligned}
 & \underbrace{(1 - \delta)[(\delta - w)^2 - w^2]}_{\text{shifted interior actions}} + \underbrace{\int_{1-\delta}^1 [(\bar{a} - \theta - w)^2 - w^2] d\theta}_{\text{states sent to } \bar{a}} \\
 & \geq \delta \left[(1 - \delta)(\delta - 2w) + (\bar{a} - 1 - w)^2 - w^2 \right] \\
 & \geq \delta \left[(\bar{a} - 1 - \bar{w})^2 - \bar{w}^2 - 2\bar{w} \right] > 0.
 \end{aligned}$$

The first inequality uses $\bar{a} - \theta - w \geq \bar{a} - 1 - w > 0$; the second uses $(1 - \delta)(\delta - 2w) \geq -2|w| \geq -2\bar{w}$, which follows by considering $w \geq 0$ and $w < 0$, together with

$(\bar{a} - 1 - w)^2 - w^2 = (\bar{a} - 1)^2 - 2(\bar{a} - 1)w \geq (\bar{a} - 1 - \bar{w})^2 - \bar{w}^2$; the third is the choice of $\bar{a}$. Whatever the monitor gains from shifting interior actions toward $\theta + w$, it loses more from the states sent to $\bar{a}$. If $\delta > 1$, the actor takes $\bar{a}$ in every state, and the monitor’s loss is at least $(\bar{a} - 1 - \bar{w})^2 > \bar{w}^2 + 2\bar{w} \geq w^2$.

If $\hat{b} > b$, the deviation is the mirror image: with $e := \hat{b} - b \leq 1$, the actor chooses $\theta - e$ on $\theta > e$ and $1 - \bar{a}$ on $\theta < e$. Reflecting states and actions around $1/2$ maps this case to understatement with biases $(-b, -w)$, so the same bounds apply, including when $e > 1$.

Truth is therefore the monitor’s uniquely optimal choice in $\mathcal{M}^*$ for every $|w| \leq \bar{w}$, and it induces $a = \theta$ almost surely for every $b \in \text{supp}(\tau)$. This proves the claim. $\square$

◆

6 DISCUSSION

Just as how economists study and design human incentives without understanding how the brain produces each choice, we have taken a functionalist view of AI systems as agents that respond to incentives. In so doing, we have made behavior—rather than the internal workings of neural nets—the basic object of analysis. This ‘Humean’ view of AI behavior as being driven by beliefs and preferences might be viewed literally, metaphorically, or normatively.

On the literal view, modern AI systems are post-trained for goal-directed behavior (Ouyang et al., 2022; Bai et al., 2022). Reinforcement learning increases the probability of actions that receive high rewards and decreases the probability of actions that receive low rewards. Indeed, a recent turn to make AI systems economically valuable—able to perform tasks in the world—has also made them behave as optimizers. On this view, their ‘preferences’ are simply whatever has been reinforced.

On the metaphorical view, preferences need not describe any internal mental state. Instead they are, in the revealed preference tradition, nothing more than a compact account of stable choices (Gul and Pesendorfer, 2008). This reading has some empirical support. Chen et al. (2023) find that GPT choices often satisfy standard consistency tests in allocation problems. Mazeika et al. (2025) find that pairwise AI choices become more transitive and fit utility models better as models scale. On this view, our conception of AI systems as having preferences are simply a useful analytical tool to characterize how AI systems behave.³⁰

On the normative view, preference maximization is a benchmark for coherent agency. We should expect strong AI systems that are deployed in markets and compete

³⁰Measured preferences can depend on the prompt, so our results can be interpreted as conditional on the prompt.

against each other to exhibit coherent preferences (e.g., such as to not be money-pumped).

Future directions. We conclude with some (speculative) directions for future work.

First, our framework, while general, is ultimately static. In practice, AI agents act over long trajectories, performing complex sequences of actions that depend on each other. Of course, these trajectories of actions $(a_t)_t$ can be thought of as a single choice i.e., a member $\in A$. There is no conceptual difficulty here since we can simply set A as the space of all trajectories and, on this view, the reward $r : A \rightarrow \mathbb{R}$ is over the full path of actions.³¹ But we expect there to be analytical and computational difficulties in analyzing optimal mechanisms when the action space is large, and more structure of how time- t actions aggregate into ‘total payoffs’ (e.g., discounting the sum) can lend more tractability. Perhaps more fundamentally, the designer might re-optimize the mechanism based on agents’ past actions. We think this calls for a more serious treatment of dynamic mechanism and information design.

Second, AI agents might, in practice, be uncertain about whether it is facing a test or is in deployment (what is often called ‘evaluation awareness.’) These kinds of *de se* uncertainty about where oneself is can be exploited to discipline behavior and screen misaligned agents (Chen et al., 2024). Self-locating information could even be designed to implement good behavior (Koh and Sanguanmoo, 2025). Indeed, a model’s belief about whether it is in testing or deployment depends on the realism of the evaluation (Williams et al., 2026; Krakovna et al., 2026) which suggests that the design of self-locating information is empirically relevant.

Third, we might want mechanisms to be robust to what AI agents can do (in the spirit of Carroll (2015)) and/or what information they have about the world or about each other (in the spirit of Bergemann and Morris (2005)). Such notions of robustness have been used in the context of using AI’s recommendations (Fudenberg and Liang, 2025; Dworczak and Smolin, 2026). We think the tools of robust mechanism design can be similarly brought to bear on shaping the preferences and permissions of agentic AI systems.

REFERENCES

- AMADOR, M., I. WERNING, AND G.-M. ANGELETOS (2006): “Commitment vs. flexibility,” *Econometrica*, 74, 365–396.
- ANTHROPIC (2024): “Responsible Scaling Policy (Version 2.0),” Technical Report, October 2024, <https://www.anthropic.com/responsible-scaling-policy>.
- ANTHROPIC (2025): “System Card: Claude Opus 4 & Claude Sonnet 4,” *Technical Report*.

³¹This is similar to how, in multi-stage games, we can associate a payoffs with ‘within period’ play and take the discounted sum, or associate payoffs directly with the path of all players’ actions.

- ANTHROPIC (2026a): “Claude Fable 5 & Claude Mythos 5 System Card,” *Technical Report*, <https://www.anthropic.com/claude-fable-5-mythos-5-system-card>.
- (2026b): “Claude Opus 5 System Card,” *Technical Report*, <https://www.anthropic.com/claude-opus-5-system-card>.
- BAI, Y., A. JONES, K. NDOUSSE, A. ASKELL, A. CHEN, N. DASSARMA, D. DRAIN, S. FORT, D. GANGULI, T. HENIGHAN, ET AL. (2022): “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback,” *arXiv preprint arXiv:2204.05862*.
- BAKER, B. ET AL. (2025): “Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation,” *arXiv preprint arXiv:2503.11926*.
- BATTAGLINI, M. (2002): “Multiple referrals and multidimensional cheap talk,” *Econometrica*, 70, 1379–1401.
- BEN-PORATH, E., E. DEKEL, AND B. L. LIPMAN (2019): “Mechanisms with Evidence: Commitment and Robustness,” *Econometrica*, 87, 529–566.
- BEN-PORATH, E. AND B. L. LIPMAN (2012): “Implementation with Partial Provability,” *Journal of Economic Theory*, 147, 1689–1724.
- BERGEMANN, D. AND S. MORRIS (2005): “Robust Mechanism Design,” *Econometrica*, 73, 1771–1813.
- (2016): “Bayes correlated equilibrium and the comparison of information structures in games,” *Theoretical Economics*, 11, 487–522.
- (2019): “Information design: A unified perspective,” *Journal of Economic Literature*, 57, 44–95.
- BOWMAN, S. R., J. HYUN, E. PEREZ, E. CHEN, C. PETTIT, S. HEINER, K. LUKOSIUTE, A. ASKELL, A. JONES, A. CHEN, ET AL. (2022): “Measuring Progress on Scalable Oversight for Large Language Models,” *arXiv preprint arXiv:2211.03540*.
- BRANDENBURGER, A. AND E. DEKEL (1993): “Hierarchies of Beliefs and Common Knowledge,” *Journal of Economic Theory*, 59, 189–198.
- CARROLL, G. (2015): “Robustness and Linear Contracts,” *American Economic Review*, 105, 536–563.
- CASPER, S., X. DAVIES, C. SHI, T. K. GILBERT, J. SCHEURER, J. RANDO, R. FREEDMAN, T. KORBAK, D. LINDNER, P. FREIRE, T. WANG, S. MARKS, C.-R. SEGERIE, M. CARROLL, A. PENG, P. CHRISTOFFERSEN, M. DAMANI, S. SLOCUM, U. ANWAR, A. SITHTHARANJAN, M. NADEAU, E. J. MICHAUD, J. PFAU, D. KRASHENINNIKOV, X. CHEN, L. LANGOSCO, P. HASE, E. BIYIK, A. DRAGAN, D. KRUEGER, D. SADIGH, AND D. HADFIELD-MENELL (2023): “Open Problems and Fundamental Limitations of Reinforcement Learning from Human Feedback,” *Transactions on Machine Learning Research*.
- CHEN, E., A. GHERSENGORIN, AND S. PETERSEN (2024): “Imperfect Recall and AI Delegation,” Working paper.
- CHEN, Y., T. X. LIU, Y. SHAN, AND S. ZHONG (2023): “The emergence of economic rationality of GPT,” *Proceedings of the National Academy of Sciences*, 120, e2316205120.
- CHRISTIANO, P., B. SHLEGERIS, AND D. AMODEI (2018): “Supervising Strong Learners by Amplifying Weak Experts,” *arXiv preprint arXiv:1810.08575*.

- COLLINA, N., S. GOEL, A. ROTH, E. RYU, AND M. SHI (2025): “Emergent Alignment via Competition,” *arXiv preprint arXiv:2509.15090*.
- CRÉMER, J. AND R. P. MCLEAN (1988): “Full Extraction of the Surplus in Bayesian and Dominant Strategy Auctions,” *Econometrica*, 56, 1247–1257.
- DENISON, C., M. MACDIARMID, F. BAREZ, D. DUVENAUD, S. KRAVEC, S. MARKS, N. SCHIEFER, R. SOKLASKI, A. TAMKIN, J. KAPLAN, ET AL. (2024): “Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models,” *arXiv preprint arXiv:2406.10162*.
- DWORCZAK, P. AND A. SMOLIN (2026): “Robust Trust,” Working paper, arXiv:2602.09490. Authors in random order.
- FIGALLI, A., Y.-H. KIM, AND R. J. MCCANN (2011): “When Is Multidimensional Screening a Convex Program?” *Journal of Economic Theory*, 146, 454–478.
- FORGES, F. (1986): “An approach to communication equilibria,” *Econometrica*, 54, 1375–1385.
- FUDENBERG, D. AND A. LIANG (2025): “Friend or foe: Delegating to an ai whose alignment is unknown,” *arXiv preprint arXiv:2509.14396*.
- GAN, T. AND Y. LI (2026): “Screening for Choice Sets,” *arXiv preprint arXiv:2601.15580*.
- GOLDOWSKY-DILL, N. ET AL. (2025): “Detecting Strategic Deception Using Linear Probes,” in *International Conference on Machine Learning*.
- GOLDSTEIN, S. AND H. LEDERMAN (2025): “What does ChatGPT want? an interpretationist guide,” *preprint*.
- GOOGLE DEEPMIND (2025): “Frontier Safety Framework (Version 2.0),” Technical Report, February 2025, <https://deepmind.google/blog/updating-the-frontier-safety-framework/>.
- GREEN, J. R. AND J.-J. LAFFONT (1986): “Partially Verifiable Information and Mechanism Design,” *Review of Economic Studies*, 53, 447–456.
- GREENBLATT, R., C. DENISON, B. WRIGHT, F. ROGER, M. MACDIARMID, S. MARKS, J. TREUTLEIN, T. BELONAX, J. CHEN, D. DUVENAUD, ET AL. (2024a): “Alignment faking in large language models,” *arXiv preprint arXiv:2412.14093*.
- GREENBLATT, R., F. ROGER, D. KRASHENINNIKOV, AND D. KRUEGER (2024b): “Stress-Testing Capability Elicitation With Password-Locked Models,” in *Advances in Neural Information Processing Systems*.
- GREENBLATT, R., B. SHLEGERIS, K. SACHAN, AND F. ROGER (2024c): “AI Control: Improving Safety Despite Intentional Subversion,” in *International Conference on Machine Learning*, PMLR, 16295–16336.
- GUL, F. AND W. PESENDORFER (2008): “The Case for Mindless Economics,” in *The Foundations of Positive and Normative Economics: A Handbook*, ed. by A. Caplin and A. Schotter, Oxford University Press, 3–39.
- HADFIELD-MENELL, D. (2021): “The Principal-Agent Alignment Problem in Artificial Intelligence,” Ph.D. thesis, University of California, Berkeley.
- HADFIELD-MENELL, D. AND G. K. HADFIELD (2019): “Incomplete contracting and AI alignment,” in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and*

- Society*, 417–422.
- HAMMOND, L., A. CHAN, J. CLIFTON, J. HOELSCHER-OBERMAIER, A. KHAN, E. MCLEAN, C. SMITH, W. BARFUSS, J. FOERSTER, T. GAVENČIAK, ET AL. (2025): “Multi-agent risks from advanced ai,” *arXiv preprint arXiv:2502.14143*.
- HART, S., I. KREMER, AND M. PERRY (2017): “Evidence Games: Truth and Commitment,” *American Economic Review*, 107, 690–713.
- HERRMANN, D., A. CHARI, I. QIAN, S. SHARVESH, AND B. A. LEVINSTEIN (2026): “A decision-theoretic approach for managing misalignment,” in *Proceedings of IASEAI Conference*, vol. 2, 217–227.
- HERRMANN, D. A. AND B. A. LEVINSTEIN (2026): “Radical AI Interpretability,” *arXiv preprint arXiv:2606.26523*.
- HOFSTÄTTER, F. ET AL. (2025): “The Elicitation Game: Evaluating Capability Elicitation Techniques,” *arXiv preprint arXiv:2502.02180*.
- HOLMSTRÖM, B. (1979): “Moral Hazard and Observability,” *The Bell Journal of Economics*, 10, 74–91.
- HUBINGER, E., C. DENISON, J. MU, M. LAMBERT, M. TONG, M. MACDIARMID, T. LANHAM, D. M. ZIEGLER, T. MAXWELL, N. CHENG, ET AL. (2024): “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training,” *arXiv preprint arXiv:2401.05566*.
- HUBINGER, E., C. VAN MERWIJK, V. MIKULIK, J. SKALSE, AND S. GARRABRANT (2019): “Risks from Learned Optimization in Advanced Machine Learning Systems,” *arXiv preprint arXiv:1906.01820*.
- HURWICZ, L. (2008): “But Who Will Guard the Guardians?” *American Economic Review*, 98, 577–585.
- HURWICZ, L. AND L. SHAPIRO (1978): “Incentive structures maximizing residual gain under incomplete information,” *The Bell Journal of Economics*, 180–191.
- JANG ET AL. (2026): “Exploration Hacking: Can LLMs Learn to Resist RL Training?” *arXiv preprint arXiv:2604.28182*.
- KENTON, Z., L. JANZER, R. GREIG, T. H. TEH, K. TYSHCHUK, J. BROWN-COHEN, H. EDWARDS, S. RAJAMANOCHARAN, N. Y. SIEGEL, N. JAQUES, AND R. SHAH (2026): “Debate Training Reduces Reward Hacking in RLAIIF,” *arXiv preprint arXiv:2608.17776*.
- KIRK, R. ET AL. (2024): “Understanding the Effects of RLHF on LLM Generalisation and Diversity,” in *International Conference on Learning Representations*.
- KOESSLER, L., J. SCHUETT, AND M. ANDERLJUNG (2024): “Risk Thresholds for Frontier AI,” *arXiv preprint arXiv:2406.14713*.
- KOH, A. (2026): “Proxy Reward Design,” *Working Paper*.
- KOH, A. AND P. MARTINEZ-BRUERA (2025): “Prices and Symmetries,” *Available at SSRN 5132200*.
- KOH, A. AND S. SANGUANMOO (2025): “Memory Correlated Equilibrium,” *MIT Working Paper*.
- KORBAK, T. ET AL. (2025): “Chain of Thought Monitorability: A New and Fragile Opportunity for AI Safety,” *arXiv preprint arXiv:2507.11473*.

- KRÄHMER, D. AND R. STRAUZ (2017): “Sequential versus Static Screening: An Equivalence Result,” *Games and Economic Behavior*, 106, 317–328.
- KRAKOVNA, V., D. LINDNER, L. HO, S. FARQUHAR, AND R. SHAH (2026): “Realistic Honeypot Evaluations for Scheming Propensity,” *arXiv preprint arXiv:2605.29729*.
- KRAKOVNA, V., J. UESATO, V. MIKULIK, M. RAHTZ, T. EVERITT, R. KUMAR, Z. KENTON, J. LEIKE, AND S. LEGG (2020): “Specification gaming: the flip side of AI ingenuity,” *DeepMind Blog*, 3.
- KWA, T., B. WEST, J. BECKER, A. DENG, K. GARCIA, M. HASIN, S. JAWHAR, M. KINNIMENT, N. RUSH, S. VON ARX, ET AL. (2025): “Measuring ai ability to complete long tasks,” *arXiv preprint arXiv:2503.14499*.
- LAINE, R. ET AL. (2024): “Me, Myself, and AI: The Situational Awareness Dataset (SAD) for LLMs,” in *Advances in Neural Information Processing Systems (Datasets and Benchmarks)*.
- LEIKE, J., D. KRUEGER, T. EVERITT, M. MARTIC, V. MAINI, AND S. LEGG (2018): “Scalable Agent Alignment via Reward Modeling: A Research Direction,” *arXiv preprint arXiv:1811.07871*.
- MAZEIKA, M., X. YIN, R. TAMIRISA, J. LIM, B. W. LEE, R. REN, L. PHAN, N. MU, A. KHOJA, O. ZHANG, ET AL. (2025): “Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs,” *arXiv preprint arXiv:2502.08640*.
- MEINKE, A., B. SCHOEN, J. SCHEURER, M. BALESI, R. SHAH, AND M. HOBBAHN (2024): “Frontier Models are Capable of In-context Scheming,” *arXiv preprint arXiv:2412.04984*.
- MERTENS, J.-F. AND S. ZAMIR (1985): “Formulation of Bayesian analysis for games with incomplete information,” *International Journal of Game Theory*, 14, 1–29.
- MILLER, N., P. RESNICK, AND R. ZECKHAUSER (2005): “Eliciting Informative Feedback: The Peer-Prediction Method,” *Management Science*, 51, 1359–1373.
- MYERSON, R. B. (1982): “Optimal Coordination Mechanisms in Generalized Principal–Agent Problems,” *Journal of Mathematical Economics*, 10, 67–81.
- NEEDHAM, J., G. EDKINS, G. PIMPALE, H. BARTSCH, AND M. HOBBAHN (2025): “Large Language Models Often Know When They Are Being Evaluated,” *arXiv preprint arXiv:2505.23836*.
- OPENAI (2024): “OpenAI o1 System Card,” *arXiv preprint arXiv:2412.16720*.
- (2025a): “GPT-5 System Card,” *Technical Report*, <https://openai.com/index/gpt-5-system-card/>.
- (2025b): “Preparedness Framework (Version 2),” *Technical Report*, April 2025, <https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbdebcd/preparedness-framework-v2.pdf>.
- OUYANG, L., J. WU, X. JIANG, D. ALMEIDA, C. WAINWRIGHT, P. MISHKIN, C. ZHANG, S. AGARWAL, K. SLAMA, A. RAY, ET AL. (2022): “Training language models to follow instructions with human feedback,” in *Advances in Neural Information Processing Systems*, vol. 35, 27730–27744.
- PEREZ, E., S. RINGER, K. LUKOŠIŪTĖ, K. NGUYEN, E. CHEN, S. HEINER, C. PETTIT, C. OLSSON, S. KUNDU, S. KADAVATH, ET AL. (2023): “Discovering Language

- Model Behaviors with Model-Written Evaluations,” in *Findings of the Association for Computational Linguistics: ACL 2023*.
- PHUONG, M. ET AL. (2024): “Evaluating Frontier Models for Dangerous Capabilities,” *arXiv preprint arXiv:2403.13793*.
- PRELEC, D. (2004): “A Bayesian Truth Serum for Subjective Data,” *Science*, 306, 462–466.
- QIU, T. A., M. CARROLL, AND C. ALLEN (2026): “Truthfulness Despite Weak Supervision: Evaluating and Training LLMs Using Peer Prediction,” in *International Conference on Learning Representations (ICLR)*.
- ROCHET, J.-C. (1987): “A Necessary and Sufficient Condition for Rationalizability in a Quasi-Linear Context,” *Journal of Mathematical Economics*, 16, 191–200.
- SAVAGE, L. J. (1954): *The Foundations of Statistics*, New York: John Wiley & Sons.
- SHAH, R., V. VARMA, R. KUMAR, M. PHUONG, V. KRAKOVNA, J. UESATO, AND Z. KENTON (2022): “Goal Misgeneralization: Why Correct Specifications Aren’t Enough for Correct Goals,” *arXiv preprint arXiv:2210.01790*.
- SHARKEY, L., B. CHUGHTAI, ET AL. (2025): “Open Problems in Mechanistic Interpretability,” *arXiv preprint arXiv:2501.16496*.
- SHEVLANE, T., S. FARQUHAR, B. GARFINKEL, M. PHUONG, J. WHITTLESTONE, J. LEUNG, D. KOKOTAJLO, N. MARCHAL, M. ANDERLJUNG, N. KOLT, ET AL. (2023): “Model Evaluation for Extreme Risks,” *arXiv preprint arXiv:2305.15324*.
- SHI, T. ET AL. (2025): “Progent: Securing AI Agents with Privilege Control,” *arXiv preprint arXiv:2504.11703*.
- SINGH, A., N. NANDA, AND S. RAJAMANOCHARAN (2026): “Why Do Models Task Game?” Blog post, AI Alignment Forum, <https://www.alignmentforum.org/posts/HACauvWhEdC6QhdS4/why-do-models-task-game>.
- TICE, C. ET AL. (2024): “Noise Injection Reveals Hidden Capabilities of Sandbagging Language Models,” *arXiv preprint arXiv:2412.01784*.
- VAN DER WEIJ, T., F. HOFSTÄTTER, O. JAFFE, S. F. BROWN, AND F. R. WARD (2025): “AI Sandbagging: Language Models Can Strategically Underperform on Evaluations,” in *International Conference on Learning Representations*.
- WEN, J. ET AL. (2024): “Adaptive Deployment of Untrusted LLMs Reduces Distributed Threats,” *arXiv preprint arXiv:2411.17693*.
- WILLIAMS, M., H. SHEAHAN, C. RAYMOND, T. KORBAK, D. PAN, P. YANG, L. MAKSHIN, N. XIE, P. GUO, I. KIVLICHAN, AND M. CARROLL (2026): “Predicting LLM Safety Before Release by Simulating Deployment,” *arXiv preprint arXiv:2607.07184*.
- XIAO AND PHUONG (2026): “Bootstrapped Monitoring: Leveraging Transparent Reasoning to Oversee Stronger AI Agents,” *arXiv preprint arXiv:2606.11998*.

A OMITTED PROOFS

Proof of Lemma 2. We first restrict the range. We then compare the two objectives on finite partitions, pass to the continuum, and choose a pointwise feasible version of the limit.

Let the least-squares loss be Q , and let the transformed loss be J :

$$Q(z) := \int [z(c) - b(c)^2]^2 dF(c), \quad J(z) := \int \left[\frac{z(c)^{3/2}}{3} - b(c)^2 \sqrt{z(c)} \right] dF(c).$$

Step 1: restrict the range. Define

$$\bar{z}(c) := (1 - \min\{c, 1/2\})^2, \quad (Tz)(c) := \min\{z(c), \bar{z}(c)\}.$$

The assumptions on bias imply

$$b(c)^2 \leq \begin{cases} c^2 \leq (1 - c)^2, & c \leq 1/2, \\ 1/4, & c \geq 1/2, \end{cases} = \bar{z}(c),$$

while $(1 - c)^2 \leq \bar{z}(c) \leq 1$. The minimum of two weakly decreasing functions is weakly decreasing, so Tz is feasible whenever z is feasible in problem (IRON). On the set where $z > \bar{z}$,

$$[z - b(c)^2]^2 - [\bar{z} - b(c)^2]^2 = [z - \bar{z}] [z + \bar{z} - 2b(c)^2] > 0.$$

Thus truncation weakly lowers Q , and lowers it strictly if it changes z on a set of positive F -probability. We can restrict attention to

$$\mathcal{Z} := \left\{ z : \begin{array}{l} z \text{ is measurable and pointwise weakly decreasing,} \\ (1 - c)^2 \leq z(c) \leq \bar{z}(c) \end{array} \right\}.$$

For every $z \in \mathcal{Z}$,

$$\begin{aligned} z(c) &= (1 - c)^2 \quad \text{for } c \leq 1/2, \\ (1 - c)^2 &\leq z(c) \leq \frac{1}{4} \quad \text{for } c \geq 1/2. \end{aligned}$$

Step 2: compare the finite problems. If $\bar{c} \leq 1/2$, the bounds fix $z(c) = (1 - c)^2$, so the result follows. Suppose instead that $\bar{c} > 1/2$. Take a finite ordered partition $(I_k)_{k=1}^m$ of

$$[\max\{\underline{c}, 1/2\}, \bar{c}]$$

into intervals of positive F -probability. For each cell, define

$$w_k := F(I_k), \quad y_k := \mathbb{E}[b(c)^2 \mid c \in I_k], \quad \ell_k := \sup_{c \in I_k} (1 - c)^2.$$

The feasible cell values form the compact convex set

$$\mathcal{Z}_m := \left\{ (z_1, \dots, z_m) : z_1 \geq \dots \geq z_m, \quad \ell_k \leq z_k \leq \frac{1}{4} \text{ for every } k \right\}.$$

Each ℓ_k is positive. Up to a term in Q that does not depend on the cell values, the two objectives for cell-constant functions are

$$Q_m(z) := \sum_{k=1}^m w_k (z_k - y_k)^2, \quad J_m(z) := \sum_{k=1}^m w_k \left[\frac{z_k^{3/2}}{3} - y_k \sqrt{z_k} \right].$$

For every feasible $z_k > 0$,

$$\frac{\partial J_m}{\partial z_k} = \frac{1}{4\sqrt{z_k}} \frac{\partial Q_m}{\partial z_k}, \quad \frac{\partial^2 J_m}{\partial z_k^2} = \frac{w_k(z_k + y_k)}{4z_k^{3/2}} > 0.$$

The positive weights make Q_m strictly convex. Let $z^m = (z_1, \dots, z_m)$ be its unique minimizer on $\mathcal{Z}_m$. Because the feasible set is a convex polyhedron, its first-order conditions give nonnegative multipliers α_k , λ_k , and μ_k for the order, lower-bound, and upper-bound constraints. With $\alpha_0 = \alpha_m = 0$, stationarity requires

$$\frac{\partial Q_m}{\partial z_k} + \alpha_{k-1} - \alpha_k - \lambda_k + \mu_k = 0 \quad (k = 1, \dots, m),$$

with the corresponding complementary-slackness conditions. Set

$$s_k := \frac{1}{4\sqrt{z_k}}, \quad \tilde{\alpha}_k := s_k \alpha_k, \quad \tilde{\lambda}_k := s_k \lambda_k, \quad \tilde{\mu}_k := s_k \mu_k.$$

Set $\tilde{\alpha}_0 = \tilde{\alpha}_m = 0$. If $\alpha_k > 0$, complementary slackness gives $z_k = z_{k+1}$, and hence $s_k = s_{k+1}$. Multiplying each stationarity equation by s_k therefore gives

$$\begin{aligned} \frac{\partial J_m}{\partial z_k} + \tilde{\alpha}_{k-1} - \tilde{\alpha}_k - \tilde{\lambda}_k + \tilde{\mu}_k &= s_k \left[\frac{\partial Q_m}{\partial z_k} + \alpha_{k-1} - \alpha_k - \lambda_k + \mu_k \right] \\ &= 0. \end{aligned}$$

The scaled multipliers remain nonnegative and preserve complementary slackness. They therefore satisfy the first-order conditions for J_m . The displayed positive second derivative makes J_m strictly convex, so z^m is also its unique minimizer.

Step 3: pass to the continuum. Take nested interval partitions $(\mathcal{P}_m)_m$ whose mesh converges to zero, and let z_m be the unique minimizer of both finite problems. Set $z_m(c) = (1 - c)^2$ for $c \leq 1/2$. Bounded monotone functions have a pointwise convergent subsequence at every continuity point of a monotone limit. Thus, along a subsequence, z_m converges to a weakly decreasing function z° . A monotone function

has at most countably many discontinuities, so atomlessness makes this convergence F -almost sure. Boundedness then gives

$$Q(z_m) \longrightarrow Q(z^\circ), \quad J(z_m) \longrightarrow J(z^\circ),$$

where we relabel the subsequence.

Fix any $z \in \mathcal{Z}$. For the cell $I_m(c) \in \mathcal{P}_m$ that contains $c > 1/2$, define the upper step approximation

$$z^{(m)}(c) := \begin{cases} (1-c)^2, & c \leq 1/2, \\ \sup_{x \in I_m(c)} z(x), & c > 1/2. \end{cases}$$

This function belongs to the finite feasible set and converges to z at every continuity point of z , hence F -almost surely. Finite optimality and dominated convergence imply

$$Q(z^\circ) \leq Q(z), \quad J(z^\circ) \leq J(z).$$

Thus z° minimizes both objectives over $\mathcal{Z}$, and its bounds hold F -almost surely.

Step 4: construct a pointwise solution. Define

$$\hat{z}(c) := \begin{cases} (1-c)^2, & c \leq 1/2, \\ \lim_{x \downarrow c} z^\circ(x), & 1/2 < c < \bar{c}, \\ \lim_{x \uparrow \bar{c}} z^\circ(x), & c = \bar{c} > 1/2. \end{cases}$$

The one-sided limits exist, and $\hat{z}$ differs from z° only at countably many points. It is weakly decreasing on each side of $1/2$. If $1/2$ lies in the capability interval, the almost-sure upper bound and full support imply

$$\lim_{x \downarrow 1/2} z^\circ(x) \leq \frac{1}{4} = \hat{z}(1/2),$$

so $\hat{z}$ is weakly decreasing on the full interval.

Suppose that either bound fails at some c_0 . One-sided continuity of $\hat{z}$ and continuity of the bounds then give a relative one-sided interval on which the same bound fails. Full support gives this interval positive probability, contradicting the almost-sure bounds. Therefore

$$(1-c)^2 \leq \hat{z}(c) \leq (1 - \min\{c, 1/2\})^2 \quad \text{for every } c.$$

Hence $\hat{z} \in \mathcal{Z}$. Since $\hat{z} = z^\circ$ F -almost surely, it minimizes both objectives.

Finally, Q is strictly convex. The transformed integrand is also strictly convex wherever $z > 0$, since

$$\frac{\partial^2}{\partial z^2} \left[\frac{z^{3/2}}{3} - b(c)^2 \sqrt{z} \right] = \frac{1}{4\sqrt{z}} + \frac{b(c)^2}{4z^{3/2}} > 0.$$

The lower bound gives $z(c) > 0$ for every $c < 1$, and atomlessness gives $F(\{1\}) = 0$. Each objective therefore has at most one minimizer up to F -null sets. Step 1 shows that every minimizer of problem (IRON) lies in $\mathcal{Z}$ up to a null set. Setting $z^* := \hat{z}$ proves all the claims. $\square$

Proof of Proposition 4. We first reduce the mechanism to one cap. We then optimize that cap.

Step 1: reduce the mechanism to one cap. Fix any incentive-compatible mechanism and let D_b be the nonempty closed permission set assigned to report b . Because D_b is nonempty and closed, every target has a nearest point in D_b . Condition (INT) implies

$$0 \leq b \leq 1 - \sqrt{\mathbb{E}[b^2]} \leq 1 \quad \tau\text{-almost surely,}$$

so every bias moment used below is finite. We now prove directly that one cap weakly dominates the assigned permission sets. For a closed set D , let $d_D(x)$ be the distance from x to D . Define the agent's loss and the human's loss when the bias is t by

$$A_D(t) := \int_t^{t+1} d_D(x)^2 dx, \quad H_D(t) := \int_0^1 (a_D(\theta, t) - \theta)^2 d\theta,$$

where $a_D(\theta, t)$ is a nearest point in D to $\theta + t$. Ties occur only at gap midpoints and do not affect either integral.

We first need one distance inequality. If d is a nonnegative, one-Lipschitz function on $[0, 1]$ and $y = \int_0^1 d(x)^2 dx$, then

$$d(1)^2 - d(0)^2 \leq \Phi(y), \quad \Phi(y) := \begin{cases} (3y)^{2/3}, & 0 \leq y \leq 1/3, \\ 2\sqrt{y - 1/12}, & y \geq 1/3. \end{cases} \quad (*)$$

To prove this inequality, put $p = d(1)$. If $p \leq 1$, the Lipschitz property gives

$$y \geq \int_{1-p}^1 (p - 1 + x)^2 dx = \frac{p^3}{3}.$$

Thus $d(1)^2 - d(0)^2 \leq p^2 \leq (3y)^{2/3}$ when $y \leq 1/3$; when $y \geq 1/3$, the same difference is at most $1 \leq 2\sqrt{y - 1/12}$. If $p \geq 1$, then $d(0) \geq p - 1$ and

$$y \geq \int_0^1 (p - 1 + x)^2 dx = p^2 - p + \frac{1}{3}.$$

It follows that

$$d(1)^2 - d(0)^2 \leq p^2 - (p-1)^2 = 2p - 1 \leq 2\sqrt{y-1/12}.$$

This proves (*).

Fix a set D and a bias b . There is a unique $c \leq b+1$ such that the cap $K_c := (-\infty, c]$ gives this type the same loss:

$$A_{K_c}(b) = A_D(b).$$

Existence follows because $A_{K_c}(b)$ rises continuously from zero to infinity as c falls from $b+1$ to $-\infty$. We claim that

$$A_D(t) \leq A_{K_c}(t) \quad \text{for every } t \geq b. \quad (**)$$

Indeed, differentiation under the integral gives

$$A'_D(t) = d_D(t+1)^2 - d_D(t)^2 \leq \Phi(A_D(t)),$$

where the inequality applies (*) to the translated distance function on $[t, t+1]$. The cap attains equality because its distance function is the upper ramp $x \mapsto (x-c)_+$. Hence

$$A'_{K_c}(t) = \Phi(A_{K_c}(t)) \quad \text{for } t \geq b.$$

Let $Q(y) := \int_0^y 1/\Phi(v) dv$. This function is finite and strictly increasing. Wherever the loss is positive,

$$\frac{d}{dt}Q(A_D(t)) \leq 1 = \frac{d}{dt}Q(A_{K_c}(t)).$$

The two losses agree at $t = b$, so integration proves (**); zero-loss points follow by continuity.

The same cap also weakly lowers the human's loss for type b . The envelope formula gives

$$A'_D(b) = -2 \int_0^1 (a_D(\theta, b) - \theta - b) d\theta,$$

and therefore

$$H_D(b) = A_D(b) - bA'_D(b) + b^2.$$

At $t = b$, inequality (*) gives $A'_D(b) \leq A'_{K_c}(b)$. Since $b \geq 0$ and the two agent losses agree at b , we obtain

$$H_{K_c}(b) \leq H_D(b). \quad (***)$$

Apply this cap replacement to every assigned set D_b and write $c(b)$ for its endpoint. Take any $b < b'$ in $\text{supp}(\tau)$. Completeness lets type b' report b . Incentive compatibility and (**) therefore imply

$$A_{K_{c(b')}}(b') = A_{D_{b'}}(b') \leq A_{D_b}(b') \leq A_{K_{c(b)}}(b').$$

The loss from a cap is strictly decreasing in its endpoint below $b' + 1$. Since $c(b) \leq b + 1 < b' + 1$ and $c(b') \leq b' + 1$, this inequality requires

$$c(b) \leq c(b').$$

Thus $c(b)$ is weakly increasing and hence measurable on $\text{supp}(\tau)$. Meanwhile, $(***)$ shows that each loss-matched cap weakly lowers the human's loss for its matched type. We use these caps only as comparison objects; their report-dependent profile need not preserve every incentive constraint.

It remains to pool these ordered caps. The human's loss for known bias b is minimized at

$$q(b) := \max\{1/2, 1 - b\},$$

as the derivative calculated in Step 2 below confirms. The function $q(b)$ is weakly decreasing, while $c(b)$ is weakly increasing. For each b , consider the closed interval between $c(b)$ and $q(b)$. If $b < b'$, then

$$\min\{c(b), q(b)\} \leq c(b') \leq \max\{c(b'), q(b')\}$$

and

$$\min\{c(b'), q(b')\} \leq q(b) \leq \max\{c(b), q(b)\}.$$

These inequalities show that the lower endpoint of any such interval is no greater than the upper endpoint of any other. Taking the supremum of the lower endpoints and the infimum of the upper endpoints gives

$$\sup_{b \in \text{supp}(\tau)} \min\{c(b), q(b)\} \leq \inf_{b \in \text{supp}(\tau)} \max\{c(b), q(b)\}.$$

Both bounds are finite because $1/2 \leq q(b) \leq 1$. Choose a common endpoint c between them. It lies between $c(b)$ and $q(b)$ for every supported bias. Moving a cap toward $q(b)$ weakly lowers the human's loss, so $(***)$ gives

$$H_{K_c}(b) \leq H_{K_{c(b)}}(b) \leq H_{D_b}(b) \quad \text{for every } b \in \text{supp}(\tau).$$

Assigning K_c after every report is incentive compatible. We can therefore restrict attention to one cap $(-\infty, \bar{a}]$.

Step 2: optimize the cap. Let $L_b(\bar{a})$ denote the human's expected loss for type b . For $\bar{a} \leq 1$, direct integration and differentiation give

$$\frac{dL_b(\bar{a})}{d\bar{a}} = \begin{cases} 2\bar{a} - 1, & \bar{a} \leq b, \\ b^2 - (1 - \bar{a})^2, & b < \bar{a}. \end{cases}$$

Condition (INT) bounds b by one, so dominated convergence justifies taking expectations of these derivatives. Set $\bar{a}^* = 1 - \sqrt{\mathbb{E}_{b \sim \tau}[b^2]}$. Condition (INT) gives $b \leq \bar{a}^* \leq 1$

almost surely. If $\bar{a} < 0$, then $\frac{dL_b(\bar{a})}{d\bar{a}} = 2\bar{a} - 1 < 0$ for every type. If $0 \leq \bar{a} < \bar{a}^*$, then

$$\mathbb{E}_{b \sim \tau} \left[\frac{dL_b(\bar{a})}{d\bar{a}} \right] \leq \mathbb{E}_{b \sim \tau}[b^2] - (1 - \bar{a})^2 < 0,$$

because $2\bar{a} - 1 \leq b^2 - (1 - \bar{a})^2$ whenever $b \geq \bar{a}$. If $\bar{a}^* < \bar{a} \leq 1$, interiority gives $b < \bar{a}$ almost surely, so

$$\mathbb{E}_{b \sim \tau} \left[\frac{dL_b(\bar{a})}{d\bar{a}} \right] = \mathbb{E}_{b \sim \tau}[b^2] - (1 - \bar{a})^2 > 0.$$

For $\bar{a} \geq 1$, a binding cap has derivative $b^2 - (\bar{a} - 1)^2 > 0$, while a cap that does not bind has derivative zero. Loss is therefore nondecreasing above one. These inequalities prove part (i).

It remains to calculate the value. Condition (INT) gives $0 \leq b \leq \bar{a}^* \leq 1$. More generally, whenever $b \leq \bar{a} \leq 1$ almost surely,

$$L(\bar{a}) = \mathbb{E}_{b \sim \tau} \left[\int_0^{\bar{a}-b} b^2 d\theta + \int_{\bar{a}-b}^1 (\bar{a} - \theta)^2 d\theta \right] = \mathbb{E}_{b \sim \tau} \left[\bar{a}b^2 - \frac{2}{3}b^3 + \frac{(1 - \bar{a})^3}{3} \right].$$

Substituting $\bar{a}^* = 1 - \sqrt{\bar{b}^2 + \sigma^2}$ gives part (ii). Assigning D^* after every report is feasible and incentive compatible, so the bound is attained. $\square$