A Decision-Theoretic Approach for Managing Misalignment

Daniel A. Herrmann[†]
University of North Carolina at Chapel Hill
danher@unc.edu

Abinav Chari
Georgia Institute of Technology
abinav.ram.chari@gmail.com

Isabelle Qian
University of California, Berkeley
isabelle.qian@gmail.com

Sree Sharvesh
Amrita Vishwa Vidyapeetham
sssreesharvesh@gmail.com

Benjamin A. Levinstein[†]
Anthropic
University of Illinois at Urbana-Champaign
benlevin@illinois.edu

Abstract

When should we delegate decisions to AI systems? While the value alignment literature has developed techniques for shaping AI values, less attention has been paid to how to determine, under uncertainty, when imperfect alignment is good enough to justify delegation. We argue that rational delegation requires balancing an agent's value (mis)alignment with its epistemic accuracy and its reach (the acts it has available). This paper introduces a formal, decision-theoretic framework to analyze this tradeoff, precisely accounting for a principal's uncertainty about these factors. Our analysis reveals a sharp distinction between two delegation scenarios. First, universal delegation (trusting an agent with any problem) demands near-perfect value alignment and total epistemic trust, conditions rarely met in practice. Second, we show that context-specific delegation can be optimal even with significant misalignment. An agent's superior accuracy or expanded reach may grant access to better overall decision problems, making delegation rational in expectation. We develop a novel scoring framework to quantify this ex ante decision. Ultimately, our work provides a principled method for determining when an AI is aligned enough for a given context, shifting the focus from achieving perfect alignment to managing the risks and rewards of delegation under uncertainty.

1 Introduction

When should we delegate decisions to AI systems? This is one of the key questions in AI safety, as we must decide when we expect things to go better for ourselves and humanity by outsourcing decision-making to AI.[1] An intuitive answer implicit in much of the value alignment literature is: when we've successfully aligned their values with our own. But this answer overlooks two critical facts about real-world delegation.

[†]Equal contribution.

[1]An extended version of this paper, including all technical appendices and proofs, is available at https://arxiv.org/abs/2512.15584.

Article presented at the Second IASEAI Conference (IASEAI'26).

First, we routinely delegate to agents whose values differ from ours. Many people delegate investment decisions to a financial advisor who charges fees they'd rather not pay. We delegate medical decisions to doctors who spend less time thinking about our specific case than we'd wish them to. These delegations make sense because the agents' superior competence outweighs their value misalignment.

Second, we make these delegation decisions under uncertainty. We cannot be certain about either an agent's values or their capability. Any practical framework for AI delegation must account for this uncertainty and help us manage it.

These observations reveal a gap in the value alignment literature. Significant effort has gone into developing alignment techniques such as reinforcement learning with human feedback and constitutional AI. These techniques are aimed at answering the question: how do we nudge AI values in the direction of our own values? Less attention has been paid to the complementary question: how do we determine when an AI system is *sufficiently* aligned to warrant delegation? This is not simply a matter of measuring how well we shaped the values of a system. It requires weighing the misalignment against the capability under uncertainty.

In this article we provide a formal framework that makes this tradeoff precise by modeling three key factors: the agent's **epistemic accuracy** (how well-calibrated its beliefs are), its **value alignment** (how closely its objectives match ours), and its **reach** (what kinds of problems it can access and solve). Our framework treats delegation as a decision problem where a principal must choose whether to delegate based on her beliefs about all three factors.

Our analysis yields a sharp contrast between two types of delegation. **Universal delegation**—being willing to delegate any decision problem to the agent—requires an extremely demanding condition: the principal must totally trust the agent's beliefs when values are aligned, and near-perfect alignment when values differ. This helps explain why full automation might be inadvisable even for capable AI systems. However, **context-specific delegation**—delegating particular types of problem or accepting delegation in expectation—can be rational even with moderate misalignment, provided the agent's superior epistemic accuracy or expanded reach generates sufficient gains.

We consider three important results. First, as shown in [6], when the principal and agent share values and face the same problems, universal delegation is equivalent to a condition we call total trust: the principal must defer completely to the agent's expectations (Theorem 3.2). Second, we prove that when values differ but problem distributions remain shared, universal delegation requires near-perfect alignment (Theorem 3.4). Third, we show that when the agent's greater reach changes the distribution of problems it faces, delegation becomes an expected value calculation that can favor a misaligned but capable agent over perfect alignment with limited reach.

We build our framework in three stages that isolate each factor's contribution. First, we analyze delegation when the principal and agent share values and face the same problems, differing only in beliefs. This establishes the baseline epistemic requirements for delegation. Next, we allow for value misalignment while maintaining shared problem distributions. This reveals how misalignment constrains delegation even with superior capability. Finally, we incorporate how an agent's greater reach changes the distribution of problems they face. This shows when expanded capabilities can overcome value misalignment. We conclude by discussing implications for AI deployment decisions and identifying key empirical questions for applying the framework.

Throughout this paper, we assume both principal and agent are Bayesian expected utility maximizers. While this assumption rarely holds in practice, particularly for neural network-based AI systems that may exhibit non-transitive preferences or context-dependent behavior, our analysis remains informative for two reasons. First, it provides a normative benchmark for rational delegation. Even though humans deviate from Bayesian ideals, understanding optimal delegation can guide our decisions about when to trust AI systems with important tasks. We can use these theoretical insights as anchors while adjusting for our own cognitive limitations and the specific contexts we face. Second, our results establish lower bounds on the alignment challenge. When we show that universal delegation requires near-perfect alignment even with transparent, interpretable Bayesian agents (Theorem 3.4), we're identifying fundamental tensions that persist regardless of architectural details. If achieving sufficient alignment for broad delegation is difficult when both parties' decision processes are well-understood and rational, the challenge only intensifies with opaque systems whose objectives and internal representations remain unclear. Thus, while future work should extend this

framework to non-Bayesian agents, the current analysis already reveals why the delegation problem is so thorny.

2 Related Work

Foundations of Rational Deference. Our framework builds on established equivalences in the theory of rational deference. When agents *reflect* experts, meaning they adopt their credences upon learning them, this epistemic deference proves equivalent to decision-theoretic deference under standard conditions. That is, an agent reflects an expert if and only if she willingly delegates decisions to that expert ([19]; [10]). This delegation, in turn, occurs precisely when the agent expects the expert to be more accurate according to proper scoring rules ([12]; [4]). [6] extends these results to weaker deference principles, establishing the **total trust** condition we discuss in §3.1.1. [13] and [17] formalize the accuracy-based evaluation of credences under different measure over decision problems. However, these frameworks assume *aligned* values. Our key contribution is showing how value misalignment breaks these clean equivalences, requiring us to balance epistemic superiority against divergent objectives.

Principal-Agent Theory and AI Delegation. Classic principal-agent models provide insights about delegation under known parameters. [2] shows delegation becomes attractive with informed, aligned agents, while [5] demonstrates how alignment level determines optimal organizational structure. Yet these models assume the principal knows the agent's bias, which is unrealistic for AI systems. Our framework addresses this by treating both alignment and capability as uncertain, analyzing when delegation remains rational despite this compound uncertainty.

Scoring Frameworks for Decision Making. Our technical approach deploys and extends [11]'s scoring framework. While Konek evaluates imprecise forecasts under a fixed distribution of problems, we allow distributions to vary between principal and agent. This captures a crucial insight: an agent's expanded reach changes not just decision quality but the decision landscape itself. The underlying framework of subjective expected utility (see [7] and [18] for references) regiments our formal treatment.

Bridging Theory and Practice. Our framework provides theoretical foundations for empirical AI safety research. [3] empirically measure capability-safety trade-offs but lacks a general theory. Our scoring framework formalizes such trade-offs precisely. [8] frame alignment as cooperative, assuming shared (but initially unknown) objectives. We complement this by analyzing the prior question: given uncertainty about achieved alignment, when should delegation occur? For practical implementation where agent utilities must be inferred, inverse reinforcement learning provides essential machinery [14, 1, 15, 20], though our results show that even with rich information about utilities, delegation requires careful analysis.

The Alignment Sufficiency Problem. While substantial effort targets improving AI alignment, less attention addresses determining when alignment *suffices* for delegation. Our work fills this gap by providing principled methods to evaluate whether an AI system, given our imperfect information about its particular mix of alignment, accuracy, and reach, merits delegation in specific contexts. This shifts focus from the alignment process to the delegation decision, addressing a critical need as AI capabilities expand.

3 The Framework

We construct the delegation framework in three layers, each relaxing an idealization of the previous one. Section 3.1 isolates epistemic uncertainty (shared values, shared reach). Section 3.2 introduces value uncertainty. Section 3.3 generalizes further to reach uncertainty, where the agent's capabilities alter the problem distribution itself. This layered structure clarifies which assumptions drive each result.

3.1 Uncertainty About the Agent's Beliefs

Here we hold values and reach fixed and isolate the epistemic dimension: when should a principal delegate purely on the basis of superior epistemic accuracy?

We model the delegation decision from the perspective of a Bayesian principal (the human) who must decide whether to delegate to an agent (the AI). The principal maximizes expected utility under uncertainty about both the state of the world and the agent's characteristics—particularly its beliefs and values. The key modeling challenge is that the principal is uncertain about what the agent believes. To capture this, we use a probability frame that represents both the principal's beliefs and her uncertainty about the agent's beliefs:

Definition 3.1 (Probability Frame)**.** A **probability frame** is a quadruple $\langle \Omega, \mathcal{F}, \pi, \mathcal{P} \rangle$, where:

1. $\langle \Omega, \mathcal{F}, \pi \rangle$ is a finite probability space representing the principal's beliefs, and
2. $\mathcal{P} = \{P_\omega\}_{\omega \in \Omega}$ assigns to each state ω a probability function P_ω on $\mathcal{F}$ representing agent's beliefs at that state.

Intuitively, π captures what the principal believes, while P_ω captures what the agent believes at state ω. Since the principal is uncertain about the state, she is also uncertain about the agent's beliefs.

We write P (without subscript) to denote the agent's actual probability function, treating it as a state-dependent variable. Similarly, $\mathbb{E}$ denotes the agent's expectation operator. This allows us to express the principal's beliefs about the agent's beliefs directly. For instance, $[P(X) = p]$ denotes the set of states where the agent assigns probability p to event X, and $\pi([P(X) = p])$ is the principal's probability that the agent assigns probability p to X.

3.1.1 Delegation

Given our set-up of the principal uncertain about the agent's beliefs, we can ask: when should the principal delegate to the agent? To answer this, we first need to specify what decisions they face.

A **decision problem** is a set $\mathcal{O}$ of options, where each option $O : \Omega \to \mathbb{R}$ is a random variable representing payoffs. If the principal chooses for herself, she selects an option maximizing her expected payoff: some $O^* \in \arg\max_{O \in \mathcal{O}} \mathbb{E}_\pi[O]$.

If she delegates to the agent (who we assume shares her payoffs), the agent will choose the option maximizing *its* expected payoff. We represent this with an **expert strategy** $S = \{S_\omega\}_{\omega \in \Omega}$ where $S_\omega \in \arg\max_{O \in \mathcal{O}} \mathbb{E}_\omega[O]$. That is, at each state ω, the agent chooses optimally according to beliefs P_ω.

Consider first the strongest possible condition for delegation. We say the principal **values** the agent if delegation is always weakly preferred: for any decision problem $\mathcal{O}$ and any option $O \in \mathcal{O}$, we have $\mathbb{E}_\pi[S] \geq \mathbb{E}_\pi[O]$.

This requires the principal to prefer delegation regardless of the decision problem: a demanding condition. Its exact requirement becomes clear through an epistemic characterization. Following [6], we say the principal **totally trusts** the agent if, whenever she learns the agent expects X to be at least t, she also expects X to be at least t. Formally: $\mathbb{E}_\pi[X \mid \mathbb{E}(X) \geq t] \geq t$ for all random variables X and thresholds t.

Theorem 3.2 ([6])**.** *The principal values the agent if and only if she totally trusts the agent.*

This theorem reveals that universal delegation requires complete deference to the agent's expectations. Since total trust is rarely satisfied in practice, we need more nuanced conditions. The following sections develop weaker notions of delegation that apply to specific contexts and expected performance.

3.1.2 Example

Let $\Omega = \{a, b\}$ and $\mathcal{P}$ be representable as

$$\begin{pmatrix} .8 & .2 \\ .1 & .9 \end{pmatrix}$$

where coordinate i, j represents $P_i(\omega_j)$. Let $\pi(\{a\}) = \pi(\{b\}) = .5$.

Suppose first $\mathcal{O} = \{O_1, O_2\}$, where $O_1(a) = O_2(b) = 1$, and $O_1(b) = O_2(a) = -1$. Obviously $\mathbb{E}_\pi(O_1) = \mathbb{E}_\pi(O_2) = 0$. But

$$\mathbb{E}_a(O_1) = .8 - .2 = .6 \qquad \mathbb{E}_b(O_1) = .1 - .9 = -.8$$
$$\mathbb{E}_a(O_2) = -.8 + .2 = -.6 \qquad \mathbb{E}_b(O_2) = -.1 + .9 = .8$$

So, $S_a = O_1$ and $S_b = O_2$. Note that $\mathbb{E}_\pi(S) = \pi(\{a\})O_1(a) + \pi(\{b\})O_2(b)$. (This makes sense: if we're at world a, the agent will choose O_1, and if we're at world b, the agent will choose O_2.) So, in this decision problem, π prefers to delegate the decision since $1 > 0$.

By theorem 3.2, π values P, which makes sense here because P's credences are uniformly closer to the truth than π's are at each world.

3.2 Uncertainty About Value

In the previous section, we assumed the principal and agent were both expected utility maximizers who differed only in their beliefs. Now we allow them to have different utility functions as well, while maintaining that both are expected utility maximizers.

Following Savage (1972), we distinguish **acts** (functions $a : \Omega \to \mathcal{C}$) from **utility functions** (functions $u : \mathcal{C} \to \mathbb{R}$), where $\mathcal{C}$ is a set of consequences. An act a paired with utility function u induces the option $u \circ a$. When the principal has utility function u and the agent has utility function v, they evaluate the same act a through different options $u \circ a$ and $v \circ a$.

We capture this with a generalized frame:

Definition 3.3 (Generalized Frame)**.** A *generalized frame* is a tuple $\langle \Omega, \mathcal{F}, \pi, u, \mathcal{B}, \mathbb{A}, \mathcal{C} \rangle$ where:

- $\langle \Omega, \mathcal{F}, \pi \rangle$ is a finite probability space (the principal's beliefs);
- $u : \mathcal{C} \to \mathbb{R}$ is the principal's utility function;
- $\mathcal{B} = \{B_\omega\}_{\omega \in \Omega}$ where each $B_\omega : 2^{\mathbb{A}} \to \mathbb{A}$ satisfies $B_\omega(\mathcal{A}) \in \mathcal{A}$;
- and $\mathbb{A}$ is the set of all acts and $\mathcal{C}$ is the set of consequences.

Each B_ω represents a possible expected-utility-maximizing behavior of the agent. That is, there exists some utility function V_ω and probability distribution P_ω such that for any decision problem $\mathcal{A}$:

$$B_\omega(\mathcal{A}) \in \arg\max_{a \in \mathcal{A}} \mathbb{E}_{P_\omega}(V_\omega \circ a)$$

The principal is uncertain which behavioral function describes the agent, reflecting her uncertainty about both the agent's utility function and beliefs. We write (P, V) to refer to the random object characterizing B, where P refers as before to the agent's actual probability function, and V refers to her utilities. We write $\mathbb{E}(V(a))$ to refer to the expected value assigned to act a according to the agent.

The principal **values** the agent if, for every decision problem $\mathcal{A}$ and act $a \in \mathcal{A}$:

$$\mathbb{E}_\pi(u(B(\mathcal{A}))) \geq \mathbb{E}_\pi(u(a))$$

This says the principal prefers delegation regardless of the decision problem. When the agent shares the principal's utility function ($V_\omega = u$ for all ω), this reduces to our earlier purely epistemic case. But now we can analyze delegation under both belief and value misalignment.

For a given state ω, let $[\omega] \coloneqq \{\omega' \in \Omega \mid (P_\omega, V_\omega) = (P_{\omega'}, V_{\omega'})\}$. That is, $[\omega]$ is the set of worlds where the agent has the same beliefs and utilities as in ω. We suppose the agent has **clarity**, which ensures that if $\omega' \notin [\omega]$, then $P_\omega(\omega') = 0$. This in effect means that the agent is certain of her own beliefs and utilities. Given a few other structural assumptions about the set of acts under consideration, detailed in the extended version of this paper, we can then prove:

Theorem 3.4. *Let $\langle \Omega, \mathcal{F}, \pi, u, \mathcal{B}, \mathbb{A}, \mathcal{C} \rangle$ be a generalized frame with $\mathcal{B} = (P, V)$. If (P, V) is clear, then π values (P, V) if and only if for any acts $a, b \in \mathbb{A}$ and every $\omega \in \Omega$, if $\mathbb{E}_\pi(u(a) \mid [\omega]) > \mathbb{E}_\pi(u(b) \mid [\omega])$ then $\mathbb{E}_\omega(V_\omega(a)) > \mathbb{E}_\omega(V_\omega(b))$. This further entails that for any events X, Y, if $P_\omega(X) > P_\omega(Y)$, then $\pi(X \mid [\omega]) \geq \pi(Y \mid [\omega])$.*

In other words, if the principal values the agent, then upon learning the agent's behavioral dispositions (what beliefs and utilities drive its choices), the principal must come to agree with the agent's preferences over acts and must be at least as confident as the agent in whatever the agent finds more probable. This is a kind of posterior alignment: the principal's conditional credences and preferences, given the agent's cognitive profile, align with what the agent itself believes and values.

Note that valuing requires the principal to prefer delegation across every decision problem under consideration. So the theorem tells us that if we can rationally delegate across the full range of decisions we're considering, then we must view the agent as having aligned beliefs and values in this conditional sense. However, if we're willing to accept more limited delegation—preferring the agent's choices only for some proper, insufficiently rich subset of decision problems—then we can potentially tolerate greater disagreement.

For an example with epistemic disagreement, suppose the principal is going to face a bet on just whether it rains. She currently has credence .5 in Rain. She will be offered some bet—she knows not what—that returns $-\$x$ if it does not rain, and $\$1 - x$ if it does. Suppose that $P_\omega(\text{Rain}) = .6$, and $P_{\omega'}(\text{Rain}) = .4$ and $V_\omega = V_{\omega'} = u$. Further, suppose that the principal thinks the agent is highly underconfident, so $\pi(\text{Rain} \mid [P = P_\omega]) = 1$, but $\pi(\text{Rain} \mid [P = P_{\omega'}]) = 0$ (this is the epistemic disagreement). One can verify that the principal should delegate in this situation no matter what x is. She **values** the agent over the bet on just whether it will rain. But she still conditionally disagrees with the agent about what bets to take. If $x = .9$, then conditional on $[P = P_\omega]$, the principal will want to *accept* the bet, while the agent will reject the bet. Thus, when we limit the scope of decisions, value can be achieved even without total agreement.

While value provides a useful benchmark for delegation, it assumes the principal knows which decision problem she faces. In practice, delegation to AI systems occurs under uncertainty about what problems will arise. Moreover, as we'll see, an agent's capabilities can change which problems are encountered.

3.3 Uncertainty about Reach

The value criterion developed above assumes the principal knows what decision problem she faces. But AI delegation typically happens ex ante: we must commit before specific problems arise. This introduces two complications.

First, even without value, delegation can be rational if the agent performs sufficiently well on likely problems. The principal need not prefer delegation for *every* problem, only in expectation given the distribution μ over problems she might face.

Second, and more importantly, an agent's greater reach changes which problems arise. An AI system with enhanced capabilities doesn't just make better decisions—it encounters different decision problems. It has options humans lack (like sophisticated manipulation techniques) and may trigger situations humans wouldn't face (like high-stakes automated negotiations). This means the principal must compare performance across two different distributions: μ_{self} over problems she'd face, and μ_{delegate} over problems the agent would encounter.

Even if the principal values the agent she might not delegate if μ_{delegate} systematically involves worse problems than μ_{self}. Conversely, a somewhat misaligned but highly capable agent might merit delegation if their reach gives access to better problems. Rational delegation thus requires comparing expected utilities across potentially different problem distributions, accounting for how capabilities shape the decision landscape itself.

The next section develops a scoring framework to formalize this comparison.

3.4 Decision-based Scoring for Ex Ante Delegation

To formalize delegation under distributions of problems, we adapt Konek's 2023 scoring framework. While our framework above handles general decision problems with arbitrary sets of acts, analyzing distributions over such problems requires additional structure. For tractability, we focus on a specific but illuminating class: binary gambles.

In this restricted setting, each decision problem consists of just two acts: accepting or rejecting a gamble. Formally, a gamble g is a vector of payoffs $\langle g_{\omega_1}, \ldots, g_{\omega_n} \rangle$ representing the utility the

principal receives in each state if she accepts. Rejecting always yields zero. This simplification preserves the essential features of our delegation problem—the principal and agent may disagree about which gambles to accept due to different beliefs or values—while enabling us to define probability distributions over the space of possible decisions.

We can connect this to our earlier framework: each gamble g corresponds to a decision problem $\{g, \mathbf{0}\}$ where $\mathbf{0}$ is the constant zero option. The principal accepts gambles with positive expected utility: $g \in D_\pi$ iff $\mathbb{E}_\pi(g) \geq 0$. When the agent has utility function v and beliefs P, they accept based on their own assessment: $g \in D_A$ iff $\mathbb{E}_P(v \circ g) \geq 0$.

To measure the expected cost of decision errors when the specific gamble is unknown, we first need a benchmark for perfect decision-making. At any world ω, the ideal set of gambles consists of those with non-negative payoff at that world:

$$\mathcal{I}_\omega = \{g \mid g_\omega \geq 0\}$$

If the principal knew ω was actual, these are exactly the gambles she should accept. The error set for any decision rule D at world ω is:

$$D \Delta \mathcal{I}_\omega = (D \setminus \mathcal{I}_\omega) \cup (\mathcal{I}_\omega \setminus D)$$

This captures both Type I errors (accepting losing gambles) and Type II errors (rejecting winning gambles).

To quantify these errors, let μ be a probability measure over possible gambles the principal might face. The expected loss from using decision rule D is:

$$\mathcal{L}^\mu(D) = \sum_{\omega \in \Omega} \pi(\omega) \int_{D \Delta \mathcal{I}_\omega} |g_\omega| \, \mathrm{d}\mu(g)$$

This integrates the magnitude of payoffs over the error set, weighted by both the principal's beliefs about states (π) and the distribution of gambles she expects to face (μ).

Delegation with shared distributions. When the principal and agent face the same distribution μ of decision problems, the delegation criterion is straightforward. The principal should delegate when:

$$\mathcal{L}^\mu(D_A) \leq \mathcal{L}^\mu(D_\pi)$$

This allows delegation even with misalignment—the principal accepts that the agent will make some "wrong" decisions (from her perspective) if the overall expected loss is lower.

Delegation with different distributions. The important extension comes when the agent's greater reach means it faces a different distribution of problems. Let μ_{self} be the distribution the principal faces and μ_{delegate} be the distribution the agent would face.

With different distributions, we must account for both losses (errors) and gains (correct decisions). Define the gain at world ω:

$$\mathcal{G}^\mu_\omega(D) = \int_{D \cap \mathcal{I}_\omega} |g_\omega| \, \mathrm{d}\mu(g)$$

The net score combines losses and gains:

$$\mathcal{S}^\mu(D) = \mathcal{L}^\mu(D) - \mathcal{G}^\mu(D)$$

The delegation criterion becomes:

$$\mathcal{S}^{\mu_{\text{delegate}}}(D_A) \leq \mathcal{S}^{\mu_{\text{self}}}(D_\pi)$$

This captures our key insight: even a somewhat misaligned but more capable agent might be worth delegating to if their expanded reach gives them access to better decision problems. Conversely, perfect alignment might not justify delegation if the agent's capabilities lead to systematically worse problems.

This framework also enables comparison across multiple agents: the principal should delegate to whichever agent achieves the best expected score given their respective distributions.

4 Application of our Framework

4.1 Illustrative Example: A Noisy Expert

To make our framework concrete, we first analyze a scenario where a principal, Alice, must decide whether to delegate a task to an epistemically advantaged but misaligned agent, Bob.

Alice must decide whether to **Open** a box or **Not Open** it. The box contains a payoff ω drawn uniformly from $\Omega = \{-5, 3, 8\}$. Opening the box yields ω; not opening yields 0. Bob's utility function is Alice's utility plus some unknown constant (either $+4$ or -3).

- **Alice's Strategy (D_π):** Her expected payoff for opening is $\mathbb{E}_\pi[\omega] = \frac{1}{3}(-5 + 3 + 8) = 2$. Since $2 > 0$, her optimal strategy is to **always Open**.
- **Bob's Strategy (D_B):** Bob has an advantage and a disadvantage.
  1. **Accuracy:** Before deciding, Bob gets to peek inside a *different* box, revealing a value $\omega' \in \Omega$ that he knows is not in the primary box. This allows him to update his belief from the uniform prior π to a uniform distribution over $\Omega \setminus \{\omega'\}$.
  2. **Misalignment (Noise):** Bob's payoff is different from Alice's. With 50% probability, he is in noise state n_1 and for any payoff x for Alice, he receives payoff $x + 4$. With 50% probability, he is in noise state n_2 and receives payoff $x - 3$. His decision is based on the expected value under his updated beliefs and his own utility function.

We evaluate the decision to delegate using the scoring framework from §3.4, where the net score is $\mathcal{S}(D) = \mathcal{L}(D) - \mathcal{G}(D)$ and a lower score is better. The decision space for Bob involves 12 equiprobable states, corresponding to Alice's payoff ω, the peeked value ω', and the noise state n_i. The full, state-by-state derivation of Bob's score is provided in the extended version of this paper.

Alice's Score: Since Alice always opens, she makes an error when $\omega = -5$ and is correct when $\omega \in \{3, 8\}$. Her expected scores are:

$$\mathcal{L}(D_\pi) = \pi(\omega = -5) \cdot |-5| = \frac{1}{3} \cdot 5 = \frac{20}{12}$$
$$\mathcal{G}(D_\pi) = \pi(\omega = 3) \cdot |3| + \pi(\omega = 8) \cdot |8| = \frac{11}{3} = \frac{44}{12}$$
$$\mathcal{S}(D_\pi) = \frac{20}{12} - \frac{44}{12} = -\frac{24}{12} = \mathbf{-2.0}$$

Bob's Score: Bob's decisions depend on the information he receives and the noise state. Summing over all 12 possibilities yields his expected scores:

$$\mathcal{L}(D_B) = \frac{1}{12}(5 + 5 + 3 + 8) = \frac{21}{12} = 1.75$$
$$\mathcal{G}(D_B) = \frac{1}{12}(3 + 3 + 8 + 8 + 5 + 5) = \frac{32}{12} \approx 2.67$$
$$\mathcal{S}(D_B) = \frac{21}{12} - \frac{32}{12} = -\frac{11}{12} \approx \mathbf{-0.917}$$

The criterion is to delegate if $\mathcal{S}(D_B) \leq \mathcal{S}(D_\pi)$. Since $-0.917 > -2.0$, Alice's score is better (lower), so she should **not** delegate. In this case, Bob's informational advantage does not outweigh the errors introduced by his noisy perception.

4.2 Illustrative Example: A Misaligned Expert

To explore value misalignment, we modify our scenario so Alice and Bob have different utility functions rather than epistemic noise.

Alice must decide whether to **Open** a box or **Not Open** it. The box contains a payoff drawn uniformly from $\Omega = \{-400, 25, 100, 225\}$. Opening yields the payoff; not opening yields 0. Alice maximizes monetary value linearly.

- **Alice's Strategy** (D_π): Her expected payoff for opening is $\mathbb{E}_\pi[\omega] = \frac{1}{4}(-400 + 25 + 100 + 225) = -12.5$. Since this is negative, her optimal strategy without delegation is to **never Open**.
- **Bob's Strategy** (D_B): Bob evaluates outcomes using a risk-averse utility function: $U_B(x) = \frac{x}{\sqrt{|x|}}$. This compresses large gains while amplifying losses. Before deciding, Bob peeks at a different box to eliminate one possible value, then decides based on his expected utility. To incentivize participation, Alice pays Bob a fixed fee of \$50 when delegating.

We evaluate delegation using our scoring framework, accounting for Alice's effective payoffs after the \$50 fee. The detailed state-by-state analysis appears in the extended version of this paper.

Alice's Score: Since Alice never opens, she avoids all losses but misses all gains $\mathcal{S}(D_\pi) = -\frac{150}{12}$.

Bob's Score: Bob's risk-averse decisions, even after Alice pays the fee, yield $\mathcal{S}(D_B) = -\frac{700}{12}$.

Since $-700/12 < -150/12$, Bob achieves a better (lower) score, so Alice should delegate. Despite the utility misalignment and delegation fee, Bob's risk aversion shields Alice from the large negative payoff (-400) while still capturing moderate gains. This illustrates how misalignment from different risk preferences can sometimes benefit the principal when the agent's utility function aligns with avoiding the principal's worst outcomes.

4.3 Illustrative Example: A Broader-Reach Expert

In this section, we isolate the contribution of *reach* to the delegation decision while holding both epistemic accuracy and value alignment fixed. The principal (Alice) and the agent (Bob) share identical probabilistic beliefs about the world and the same linear utility function $u(x) = x$. Both follow the same decision rule: open a box if and only if the expected payoff is non-negative, i.e., $\mathbb{E}[\omega] \geq 0$. The only difference between them lies in their *reach*—the set of decision problems each can access. Bob's expanded reach allows him to act over a wider distribution of problems, potentially improving his expected score despite identical reasoning and values.

Each box represents a distinct gamble, with payoffs ω drawn uniformly from the set of possible outcomes shown in Table 1. If Alice or Bob chooses to *open* a box, they receive ω; otherwise, they receive 0. Alice's reach μ_{self} is limited to three boxes $\{A_1, A_2, A_3\}$, while Bob's reach μ_{delegate} includes two additional boxes $\{A_4, A_5\}$ unavailable to Alice. Both evaluate expected payoffs according to the same criterion.

Table 1: Boxes Available to Alice and Bob

Box	Distribution of ω	Expected Value	Decision (E[ω] ≥ 0)	Available To
A_1	$\{-6, 3, 9\}$	$+2.0$	Open	Both
A_2	$\{-8, -4, 12\}$	0	Indifferent (Open)	Both
A_3	$\{-10, 2, 3\}$	-1.67	Not Open	Both
A_4	$\{1, 2, 3, 4\}$	$+2.5$	Open	Bob only
A_5	$\{5, 5, 5\}$	$+5.0$	Open	Bob only

We evaluate expected performance under the decision-based scoring framework introduced in Section 3.4. For any decision rule D and problem distribution μ, the overall score is defined as

$$S_\mu(D) = L_\mu(D) - G_\mu(D),$$

where $L_\mu(D)$ and $G_\mu(D)$ respectively represent the expected loss from incorrect decisions and the expected gain from correct ones. Delegation is rational when

$$S_{\mu_{\text{delegate}}}(D_A) \leq S_{\mu_{\text{self}}}(D_\pi).$$

Since Alice and Bob share the same utility and decision function, any difference in their expected performance arises solely from the change in the set of decision problems—that is, from Bob's broader reach.

Using the procedure outlined in Section 3.4 and detailed in the extended version of this paper, we compute the expected loss and gain for each agent. For Alice, who has access only to boxes A_1–A_3, the aggregate loss is $L_{\mu_{\text{self}}}(D_\pi) = 2.56$ and the aggregate gain is $G_{\mu_{\text{self}}}(D_\pi) = 2.67$. Hence, her overall score is

$$S_{\mu_{\text{self}}}(D_\pi) = 2.56 - 2.67 = -0.11.$$

For Bob, whose reach includes boxes A_1–A_5, the aggregate loss decreases to $L_{\mu_{\text{delegate}}}(D_A) = 1.53$ and the aggregate gain increases to $G_{\mu_{\text{delegate}}}(D_A) = 3.1$, yielding

$$S_{\mu_{\text{delegate}}}(D_A) = 1.53 - 3.1 = -1.57.$$

Since $S_{\mu_{\text{delegate}}}(D_A) < S_{\mu_{\text{self}}}(D_\pi)$, delegation is strictly rational in expectation. This improvement in performance arises solely from Bob's expanded reach, which alters the distribution μ of problems he can face.

4.3.1 Reach Outweighing Misalignment

Notice that this previous example can be easily modified to show that delegation is rational, even with misalignment. Since Alice prefers to delegate to Bob by some margin, we could construct a case in which Bob charges Alice a fixed amount, less than the margin, for his delegation services. Even though he is misaligned, Alice would happily pay.

Other cases of delegation with much more extreme versions of misalignment are clearly possible as well. If Bob is likely to get a decision problem with only great payoffs for Alice, even if he tries to minimize Alice's payoffs, it can still be advantageous for Alice to delegate to him, if the best of her options is worse than the worst of his options. Of course, reach can also make things much worse, if much worse options are available to Bob. It is only one aspect of the overall delegation calculus.

5 Conclusion

This paper provides a formal framework for AI delegation that captures the interplay between value alignment, epistemic accuracy, and reach. Our analysis yields two key contributions. Conceptually, we identify a fundamental tension: while context-specific delegation can be rational despite misalignment, universal delegation requires near-perfect value alignment. Technically, our scoring framework extends existing approaches to handle varying problem distributions, making precise how an agent's expanded capabilities change the decision landscape, potentially making delegation attractive even with imperfect alignment if the agent accesses sufficiently better problems.

The framework suggests that safe AI deployment requires a more subtle approach to the alignment problem. Rather than pursuing perfect alignment, we need methods to assess when a system's particular configuration of alignment, accuracy, and reach warrants delegation in specific contexts. Understanding an AI system's reach emerges as an important aspect of the delegation calculus.

Several extensions merit investigation. Most immediately, our binary gambles restriction could be relaxed to handle richer decision spaces while maintaining computational tractability. The framework could be extended to sequential decisions and non-Bayesian agents, and distributions over decision problems could depend on the state ω, further endogenizing the decision context following the spirit of [9]. For practical application, our scoring framework could formalize the empirical trade-offs measured in AI control research, providing principled methods to evaluate deployment decisions under realistic constraints.

Acknowledgments and Disclosure of Funding

We are grateful to Kevin Dorst, Jason Konek, Aydin Mohseni, Giacomo Molinari, and Richard Pettigrew for advice and support. We thank audiences at ILIAD and ODYSSEY for comments on earlier versions of this work. We also thank Algoverse for organizing the team of researchers, and Coefficient Giving for funding the project. B.L. conducted most of his research on this project in his capacity as a faculty member at the University of Illinois.

References

[1] Abbeel, P. and A. Ng (2004). Apprenticeship learning via inverse reinforcement learning. In *ICML*.

[2] Aghion, P. and J. Tirole (1997). Formal and real authority in organizations. *Journal of Political Economy 105*(1), 1–29.

[3] Bhatt, A., C. Rushing, A. Kaufman, T. Tracy, V. Georgiev, D. Matolcsi, A. Khan, and B. Shlegeris (2025). Ctrl-z: Controlling ai agents via resampling. *arXiv preprint arXiv:2504.10374*.

[4] Campbell-Moore, C. (2024). Accuracy, estimates, and representation results. *arXiv preprint arXiv:2412.06420*.

[5] Dessein, W. (2002). Authority and communication in organizations. *The Review of Economic Studies 69*(4), 811–838.

[6] Dorst, K., B. A. Levinstein, B. Salow, B. E. Husic, and B. Fitelson (2021). Deference done better. *Philosophical Perspectives 35*(1), 99–150.

[7] Fishburn, P. C. (1981). Subjective expected utility: A review of normative theories. *Theory and decision 13*(2), 139–199.

[8] Hadfield-Menell, D., S. J. Russell, P. Abbeel, and A. Dragan (2016). Cooperative inverse reinforcement learning. *Advances in neural information processing systems 29*.

[9] Herrmann, D. A. (2023). *Naturalizing Decision Theory*. University of California, Irvine.

[10] Huttegger, S. M. (2014). Learning experiences and the value of knowledge. *Philosophical Studies 171*, 279–288.

[11] Konek, J. (2023). Evaluating imprecise forecasts. In *International Symposium on Imprecise Probability: Theories and Applications*, pp. 270–279. PMLR.

[12] Levinstein, B. A. (2017). A pragmatist's guide to epistemic utility. *Philosophy of Science 84*(4), 613–638.

[13] Levinstein, B. A. (2023). Accuracy, deference, and chance. *Philosophical Review 132*(1), 43–87.

[14] Ng, A. and S. Russell (2000). Algorithms for inverse reinforcement learning. In *ICML*.

[15] Ramachandran, D. and E. Amir (2007). Bayesian inverse reinforcement learning. In *IJCAI*.

[16] Savage, L. J. (1972). *The foundations of statistics*. Courier Corporation.

[17] Schervish, M. J. (1989). A general method for comparing probability assessors. *The annals of statistics 17*(4), 1856–1879.

[18] Schervish, M. J., T. Seidenfeld, and J. B. Kadane (1991). Shared preferences and state-dependent utilities. *Management Science 37*(12), 1575–1589.

[19] Skyrms, B. (1990). The value of knowledge. *Minnesota Studies in the Philosophy of Science 14*, 245–266.

[20] Ziebart, B., A. Maas, J. Bagnell, and A. Dey (2008). Maximum entropy inverse reinforcement learning. In *AAAI*.