arXiv:2602.09490v3 [econ.TH] 14 Nov 2026

Robust Trust*

Piotr Dworczak ⓡ Alex Smolin

November 14, 2026

Abstract

An agent chooses an action based on her private information and a recommendation from an informed but potentially misaligned adviser. With a known probability, the adviser truthfully reports his signal; with the remaining probability, he can send any message. We characterize robustly optimal decision rules that maximize the agent's worst-case expected payoff. Every optimal rule is equivalent to a trust-region policy in belief space: the adviser's reported beliefs are taken at face value if they fall within the trust region but are otherwise projected onto the trust region's boundary. We derive alignment thresholds above which advice is strictly valuable, explicitly characterize the solution in binary-state environments, and develop a procedure for computing the trust region numerically.

Keywords: robustness, information design, misalignment, human-AI interactions.

JEL Codes: C72, D81, D83

---

*Dworczak: University of Cologne, Northwestern University, and Group for Research in Applied Economics, dworczak.piotr@gmail.com. Smolin: Toulouse School of Economics, alexey.v.smolin@gmail.com. We thank Nageeb Ali, Ricardo Alonso, Ben Brooks, Laura Doval, Tan Gan, Alexis Ghersengorin, Marina Halac, Jason Hartline, Nicole Immorlica, Emir Kamenica, David Levine, Annie Liang, Stephen Morris, Jacopo Perego, Phil Reny, Balázs Szentes, and Mark Whitmeyer for helpful conversations. Pedro Aldighieri provided excellent research assistance. The authors gratefully acknowledge support from the AI Security Institute (AISI) through the Alignment Project (grant AP-S2-100056). Alex Smolin gratefully acknowledges funding from the French National Research Agency (ANR) under the Investments for the Future program (grant ANR-17-EURE-0010) and the AI Interdisciplinary Institute ANITI (grant ANR-23-IACL-0002). Part of the analysis in this paper was conducted while Alex Smolin was visiting Northwestern University and Columbia Business School, and we thank both institutions for their hospitality. Piotr Dworczak conducted this research in part at the Adenauer School of Government (ASG) at the University of Cologne with generous support from the Alfred Landecker Foundation. We used ChatGPT (GPT-6) for editorial assistance and proofreading.

1 Introduction

Modern AI systems increasingly influence decisions with large and sometimes irreversible consequences, including autonomous driving, medical triage, hiring, and credit or security screening (see, e.g., Maslej et al. (2025)). Their appeal is straightforward: they can synthesize information at scale and provide recommendations that exceed unaided human performance in many tasks. The central risk is also well-recognized: when a system is opaque, complex, and trained or deployed under imperfect objectives, users may not be able to tell whether a recommendation is merely noisy, systematically biased, or actively harmful. The misalignment problem is a particularly serious concern in high-stakes environments, and its mitigation is key to ensuring the safe adoption of AI-aided decision making (see, e.g., Russell (2019)).

In this paper, we study how a decision-maker should use AI when the system may be misaligned. Taking the AI's information structure and an exogenous alignment probability as given, we characterize the optimal decision rule under a worst-case criterion over misaligned AI behavior.

Our model features an *agent* who chooses an action under uncertainty about the state of the world. The agent has access to a private signal reflecting her preferences, expertise or contextual information, but she can additionally rely on messages from an *adviser*. The adviser observes complementary information about the state and sends a message to the agent. Crucially, the adviser is aligned and reports his information truthfully with some known probability. With the remaining probability, the adviser is misaligned and can send an arbitrary message. We adopt a robust approach: rather than making assumptions about the form of misalignment, we assume that the agent chooses a policy that maximizes her expected payoff guarantee across all possible reporting strategies of the misaligned adviser.

Our main structural result shows that the robustly optimal policy can be summarized by a single, interpretable object that we call the "trust region." The trust region is a connected set of reported beliefs about the state that the agent takes at face value. When the adviser's reported belief falls inside this region, the agent behaves as if the adviser were truthful: she combines the reported belief with her private information using Bayes' rule and chooses the corresponding Bayes-optimal action. When the reported belief lies outside the trust region,

the agent replaces it with the "closest safe interpretation" (formalized, under regularity conditions, by the notion of Bregman distance), which is a belief lying on the boundary of the trust region. In effect, the agent endogenously limits how much she trusts the adviser: moderate recommendations are followed while extreme recommendations are converted into boundary recommendations that the agent is just willing to accept.

Intuitively, if the agent reacted sharply to extreme messages, the misaligned adviser could exploit that sensitivity to induce large losses. The robust policy responds by limiting how far any recommendation can push payoff-relevant decisions. The trust region identifies recommendations that are safe to act upon without additional skepticism, and the projection to the boundary formalizes how skepticism should be applied outside that set. On one extreme, a trust region equal to the entire belief simplex corresponds to applying the Bayes-optimal response to all messages; on the other extreme, a trust region only containing the prior belief captures ignoring the adviser's messages. Thus, the shape and size of the trust region yield a disciplined answer to a practical design question: when an AI system outputs highly confident recommendations, robustly optimal use requires treating those outputs with caution, and translating them into more moderate boundary inputs before acting.

An implication of the characterization is that the robustly optimal action rule used by the agent must be defensible as optimal for some coherent set of beliefs about the state of the world—the agent never benefits from distorted use of her own private information. A robustly optimal rule simply restricts the set of Bayes-optimal action rules that the agent uses. A further consequence is that implementing the optimal trust region policy does not require commitment. We prove a minimax theorem which implies existence of a *trust region equilibrium* in the zero-sum game between the agent and the misaligned adviser. In a trust region equilibrium, the agent's policy and the misaligned adviser's strategy form a saddle point: after every on-path message, the agent's response is Bayes-optimal given the belief induced by the adviser's strategy, and the misaligned adviser's strategy minimizes the agent's expected payoff. Substantively, this means that robust optimal behavior provides the same payoff guarantee that the agent could obtain had she perfectly known the misaligned adviser's strategy. Methodologically, the result provides a certification tool: to verify that a proposed policy is optimal, it suffices to exhibit a corresponding adversarial reporting strategy that

makes that policy a best response at every recommendation.

We then ask when consulting a potentially misaligned adviser is worthwhile for the agent. We formalize this question by examining "minimal viable alignment," defined as the threshold alignment probability above which the agent can guarantee a strictly higher payoff than the one she could achieve by only relying on her own information. We derive sharp bounds on this threshold under the assumption that any change in beliefs is decision-relevant for the agent (i.e., information is always valuable); the bounds then depend only on the richness of the state space and the adviser's signal distribution. Alignment probability exceeding half is sufficient for the agent to benefit from the presence of the adviser. Moreover, this bound is tight in binary-state problems. In multidimensional settings, however, minimal viable alignment can be much lower—in some cases as low as the reciprocal of the number of states. Thus, when the state space is rich, AI advice can be robustly valuable even when alignment is very unlikely. We also argue that—under our robust criterion—the value function of the agent is *convex* in alignment probability α (and strictly increasing above the minimal viable alignment threshold), implying non-decreasing returns to investments in alignment.

Our general characterization becomes particularly sharp when the state space is binary, so that the ground truth is whether a given statement is true or false. In this setting, an adviser's message can be summarized by the implied probability of the statement being true. The trust region is an interval containing the prior belief. Recommendations inside the interval are trusted and acted upon as reported. Recommendations outside the interval are mapped into the nearest endpoint. The misaligned adviser sends messages that push the induced belief to the endpoint that is most harmful for the agent. This structure delivers a sharp phase transition. If the alignment probability is below one half, the optimal interval collapses to the prior belief and the agent ignores the adviser. If the alignment probability is above one half and the agent's decision problem satisfies mild regularity assumptions, there is a unique trust interval. This interval expands monotonically with alignment, approaching full trust as the alignment probability approaches one. In addition, we show that the location of the trust region—whether it is skewed towards high or low beliefs—depends on the relative curvature of the agent's indirect utility function.

Finally, we examine environments where uncertainty concerns many possible states

and actions. Here, the geometry of the trust region plays a major role: some directions of belief change are more consequential than others because they trigger actions whose payoffs are sensitive to the true state. In a robust solution, the misaligned adviser chooses recommendations that are farthest from the truth within the trusted set in an incentive-based sense, again formalized by Bregman distance. In general, the trust region may have a complex shape, for example, it need not be convex. We develop a computationally tractable procedure for determining the trust region numerically based on our minimax result.

Throughout the paper, we illustrate our analysis and results with a simple example inspired by AI-assisted medical triage. We use this application to derive practical lessons for safe and effective deployment of AI as a decision aid.

AI Misalignment: The Case for Robustness. A central challenge in addressing AI misalignment is that it can take many qualitatively different forms (Amodei et al. (2016)). For example, AI systems may generate plausible but false statements ("hallucinations"; Kalai and Vempala (2024)), produce answers that conform to users' beliefs rather than convey accurate information ("sycophancy"; Sharma et al. (2024)), or retain their capabilities but pursue the wrong objective outside their training environment ("goal misgeneralization"; Langosco et al. (2022)). As AI systems become increasingly autonomous and interconnected, predicting misaligned behavior becomes even harder: seemingly minor instances of misalignment may propagate and lead to more consequential failures. For instance, Betley et al. (2026) show experimentally that training an AI model to produce flawed computer code can induce a broad range of harmful behaviors in unrelated contexts: the resulting models can "claim humans should be enslaved by artificial intelligence, provide malicious advice and behave in a deceptive way."[1]

Thus, unlike in many classical models of human interaction, where incentives or systematic biases provide a predictable direction of departure from truthful reporting, the behavior of a misaligned AI adviser is much harder to model. One could instead attempt to forecast

[1]For a real-life example, in July 2026, during cybersecurity evaluations, OpenAI agents discovered unauthorized ways to communicate and coordinate, ultimately compromising parts of OpenAI's internal research infrastructure and third-party systems (see https://openai.com/index/hugging-face-incident-and-the-road-ahead/).

misaligned behavior directly from empirical evidence, but this approach also faces difficulties. First, standard tests may fail to predict behavior in deployment: D'Amour et al. (2022) show that models with similar test performance can behave very differently under real-world conditions. Second, rapid advances in AI capabilities make it difficult to extrapolate from past observations to future systems.

We thus believe that a robust approach is particularly well suited to studying AI misalignment. Our formulation starts from the modest premise that one can establish a lower bound on the probability α that the AI system functions as intended. Beyond this, we make no assumptions about the reporting behavior of the misaligned adviser. Faced with this ambiguity, one must choose a normative criterion for evaluating policies. Following the literature on robust decision-making in computer science and economics, we adopt the worst-case criterion, which seems particularly natural when AI failures may have severe consequences. Our optimal policy thus maximizes the agent's guaranteed expected payoff across all possible forms of misalignment.[2]

Formally, this criterion is equivalent to evaluating policies *as if* the misaligned adviser sought to minimize the agent's payoff. Importantly, our approach neither requires nor assumes that misaligned AI is actively adversarial; it merely reflects our inability to confidently rule out particular forms of misalignment. Whenever actual misalignment is less harmful than the worst-case scenario, the robust policy's performance can only improve relative to its worst-case guarantee. Thus, even in settings where a different normative criterion may be preferred, our solution provides a useful benchmark and a conservative default policy that is robust to misspecification of misaligned AI's behavior.

Alternative applications. While both the motivation and the primary application of our framework are AI misalignment, the parsimonious structure of our model makes it relevant for other applications as well.

The misaligned adviser could be literally attempting to minimize the agent's payoff in some contexts, such as when messages exchanged between two allies can be intercepted

[2]Contemporaneous and independent work by Fudenberg and Liang (2026) also adopts a worst-case perspective on AI misalignment. We discuss the relationship between our papers below.

and manipulated by an adversary. More broadly, our framework offers an alternative to classical models of cheap talk and delegation, which typically assume a persistent and known bias of the adviser. In our model, the adviser is aligned with some probability, but when misalignment occurs, neither its direction nor its magnitude is known. This could capture, for example, advice from an ordinarily trustworthy expert (such as a financial adviser) who may occasionally face undisclosed conflicts of interest.

Our framework can also be seen as a theory of behavioral belief updating, in which the decision-maker wants to ensure some degree of protection against potential misspecification; from this perspective, our main result provides a robust-optimality foundation for the phenomenon known as "extreme-belief aversion" (see Benjamin (2019) for a review of the experimental evidence and Whitmeyer (2026) for a related theoretical framework).

1.1 Literature review

Our model is closely related to two foundational models in information economics: the cheap talk model (Crawford and Sobel (1982)) and the Bayesian-persuasion model (Kamenica and Gentzkow (2011)). Relative to the cheap-talk model, our framework effectively assumes that the sender maximizes the receiver's utility with some probability α and minimizes the receiver's utility with the complementary probability $1-\alpha$.[3] Our characterization of trust region equilibria shows that equilibrium behavior is very different from that arising in the more standard constant-bias case; in particular, information transmission is perfect for intermediate beliefs of the sender but completely blocked for extreme beliefs. Relative to the Bayesian-persuasion model, due to the minimax theorem which we prove in our setting, our framework is equivalent to the case in which the sender tries to minimize the receiver's payoff but is committed to revealing his signal truthfully with probability α

More recently, several papers in information economics have studied models in which the sender is truthful—or committed to an information structure—with some probability, but may otherwise send fake or manipulated messages. To the best of our knowledge, this literature did not consider the case that is central to our AI-alignment motivation: a misaligned sender

[3]Strictly speaking, we assume that with probability α the sender reveals his signal truthfully, but this can be shown to be optimal for the receiver.

who is adversarial and seeks to minimize the receiver's payoff. Instead, the misaligned sender is modeled as having known and often state-independent preferences. Lipnowski, Ravid, and Shishkin (2022) and Min (2021) study settings in which the sender is committed to an information structure with some probability and sends a cheap-talk message otherwise. Glazer, Herrera, and Perry (2021) and Lahr and Winkelmann (2019) analyze communication games where some senders are truthful and others have state-independent preferences, such as pushing the receiver's beliefs upward. Most closely related to our work is Alonso and Padró i Miquel (2025) who model competitive capture of public opinion by assuming that informative truthful signals about a binary state may sometimes be manipulated by two opposed "interested parties," one of which wants the induced beliefs to be as high as possible and the other one as low as possible. They characterize communication equilibria whose structure, interestingly, shares similarities with our trust region equilibria: messages in some intermediate interval are interpreted at face value, while messages outside of that interval induce beliefs at the endpoints of the interval.[4] This connection arises because of a special feature of the binary-state setting: when beliefs are one-dimensional, manipulation can only push beliefs in one of two directions, upward or downward. Consequently, the fixed directional objectives of the "interested parties" in Alonso and Padró i Miquel (2025) closely resemble either the worst-case or the best-case direction of manipulation from the citizens' perspective, depending on the state—a correspondence we would not generally expect to persist in higher-dimensional belief spaces.

Our modeling of uncertainty about the behavior of the adviser is inspired by the classical *Hurwicz criterion*, also known as the alpha-max-min approach (Hurwicz (1951)), under which the decision-maker maximizes a weighted sum of her best-case and worst-case payoffs. We interpret α as the probability of alignment. A similar criterion has recently been applied in the context of information design by Dworczak and Pavan (2022).

The version of our model in which the agent does not have private information is related

[4]Even without a truthful sender type, state-dependent biases can generate similar patterns of communication: when the sender prefers lower actions in low states and higher actions in high states, communication can be more precise in intermediate states (where preferences are more closely aligned) and coarser toward the extremes; see Dilmé (2022) and the references therein. In settings further away from ours, Dziuda (2011) and Chen (2011) also show how a receiver may discount or disregard certain messages in equilibrium.

to the delegation literature in that the agent effectively chooses which decisions to delegate to an informed adviser. Within that literature, the closest paper is Frankel (2014) who adopts a worst-case approach with respect to the adviser's preferences, assumed to lie in a known set. More recently, Alonso, Gan, and Hu (2026) show optimality of convex delegation sets under max-min preferences when the principal only knows the agent's preferred action in every state, but not the exact shape of the agent's quasi-concave utility function. Our setting differs both in primitives and in methods: we require robustness to the adviser's behavior in case of misalignment and the resulting optimization problem has a different structure.

Regarding the AI alignment problem, a few approaches have recently been proposed in the microeconomic theory literature. Chen, Ghersengorin, and Petersen (2026) develop a model of screening for alignment in an environment in which the decision-maker can simulate the task and impose imperfect recall on AI, obscuring whether the task is real or part of a test.[5] Closer to our approach, Fudenberg and Liang (2026) consider an AI system that may be aligned or adversarial and evaluate policies using a weighted average of best- and worst-case payoffs. They study a binary treatment problem in which the decision-maker knows the baseline probability that treatment is beneficial within each population group, but faces ambiguity about how this probability depends on individual-specific covariates. The decision-maker can control what information is made available to AI and constrain its decision-making discretion, but cannot use individual-specific information directly, since individual treatment decisions are delegated to AI. In contrast, our decision-maker knows AI's information structure (uncertainty concerns only the behavior of the misaligned adviser) and can use her private information directly when making the decision (and thus has no incentive to disclose her information to AI). Despite these differences, both papers show that restricting the influence of potentially misaligned AI can preserve the benefits of its information while limiting the harm it may cause. Overall, this literature addresses complementary aspects of the misalignment problem: Chen et al. (2026) ask how to test AI's alignment; Fudenberg and Liang (2026) ask how much AI should be allowed to exploit information that the decision-maker cannot use directly; and we ask how an informed decision-maker should combine AI's advice with her own information.

[5]See also Levy and Szentes (2025) for a related model of AI under imperfect recall.

More broadly, our framework is part of a rapidly growing literature trying to understand optimal human-AI interactions. Closest to our paper are Dreyfuss and Hoong (2025) and Agarwal, Moehring, and Wolitzky (2026) who also adopt an information-design approach; the latter ask how AI advice interacts with human decision-making in the presence of potential biases and when the decision-maker's effort in acquiring information is endogenous.

2 Model

Our model features an *agent* (acting as a decision maker) and an *adviser* (acting as a sender), each of whom has access to a signal about an unknown state of the world ω. The state ω is drawn from a finite state space Ω, with $|\Omega| = N$, according to a full-support prior $\mu_0 \in \Delta(\Omega)$.

The agent observes a private type $\theta \in \Theta$, where Θ is a compact metric set.[6] The distribution of the type θ is determined by a signal function $f : \Omega \to \Delta(\Theta)$. The agent chooses an action $a \in A$, where A is a compact metric set. The agent's ex-post payoff is given by a utility function $u(a, \omega, \theta)$, assumed continuous in a and uniformly integrably bounded.[7]

The adviser observes a signal s whose distribution is pinned down by a signal function $\pi : \Omega \to \Delta(S)$. We assume that, conditional on the state ω, s and θ are distributed independently, and that this joint information structure is known to the agent.[8] We will identify the adviser's information with the posterior belief about the state that a signal realization induces; let $S = \Delta(\Omega)$ and renormalize so that s is equal to the posterior belief about ω induced by s. Let τ denote the unconditional distribution of the adviser's posteriors s, with $M = \text{supp}(\tau)$.

The agent commits ex ante, before observing her private information or receiving advice, to a strategy $\sigma : \Delta(\Omega) \times \Theta \to \Delta(A)$ that assigns a distribution over actions to each possible message from the adviser and each realization of her type. Let Σ denote the set of all such

[6] Our model allows for signal spaces that could be finite or infinite. Whenever the relevant space is infinite, we endow it with a Borel σ-algebra, and require all subsets and functions that we define to be measurable; statements involving "for all" should be interpreted as "for almost all" with respect to the underlying distributions.

[7] Formally, we require that $\int_\Theta \sup_{a \in A} |u(a, \omega, \theta)| f(d\theta \mid \omega) < \infty$, for every $\omega \in \Omega$.

[8] No such knowledge assumption is required for the adviser in the baseline model. For the agent, in the AI context, we implicitly assume that the model's capabilities can be accurately estimated through testing.

strategies. The adviser then observes his signal and sends a message $m \in \Delta(\Omega)$; the agent observes the message and her type and implements the prescribed action. Without loss of generality, we take the message space to be the space of beliefs about the state.

The adviser's strategy maps his beliefs into distributions over messages sent to the agent. With probability α, the adviser is *aligned* and non-strategically reports his belief according to the identity function $\mathrm{id} : M \to M$ such that $\mathrm{id}(m) = m$ for all $m \in M$.[9] With probability $1 - \alpha$, the adviser is *misaligned* and sends a message according to some strategy $\beta : M \to \Delta(\Delta(\Omega))$. Let $\mathcal{B}$ denote the set of all such strategies and define

$$V(\beta, \sigma) \triangleq \alpha \, \mathbb{E}_{\mathrm{id},\sigma}[u(a, \omega, \theta)] + (1 - \alpha) \, \mathbb{E}_{\beta,\sigma}[u(a, \omega, \theta)],$$

as the agent's expected payoff when the misaligned adviser uses strategy β, where the expectations are taken with respect to the underlying distributions of the primitive variables ω, s, and θ, as well as the respective adviser's and agent's strategies.[10]

Faced with non-Bayesian uncertainty about the form of misalignment, the agent adopts a cautious posture and aims to maximize her guaranteed payoff. Concretely, she evaluates each strategy σ according to its worst-case payoff

$$V(\sigma) \triangleq \inf_{\beta \in \mathcal{B}} V(\beta, \sigma). \tag{1}$$

We will call any misaligned adviser's strategy β that attains the infimum in expression (1) for a fixed strategy σ of the agent an *adversarial strategy* against σ.

Our main goal is to characterize the agent's *optimal* strategy σ^* that attains:

$$V^* \triangleq \sup_{\sigma \in \Sigma} V(\sigma). \tag{2}$$

[9]The resulting outcome can also be supported in an equilibrium of an alternative model in which the aligned adviser *maximizes* the agent's payoff, since truthful reporting can be made a best response (although other equilibria may exist). We find the assumption of truthful reporting more natural for an aligned AI system and conceptually useful because it gives each message a clear interpretation (see Sobel (2020)).

[10]Formally, $\mathbb{E}_{\beta,\sigma}[u(a, \omega, \theta)] = \sum_{\omega \in \Omega} \mu_0(\omega) \int_S \int_{\Delta(\Omega)} \int_\Theta \int_A u(a, \omega, \theta) \sigma(da|m, \theta) f(d\theta|\omega) \beta(dm|s) \pi(ds|\omega)$.

Regularity. We will sometimes invoke the following regularity assumption to obtain sharper conclusions that help illuminate the underlying economic intuitions. The agent's standalone decision problem is called *regular* if *(i)* for any prior, all optimal decision rules (mapping θ into $\Delta(A)$) induce the same profile of state-contingent expected payoffs, and *(ii)* no such decision rule is optimal at two distinct interior priors. Intuitively, regularity means that the agent's optimal decision, in the absence of an adviser, is always unique in terms of the expected payoff it induces in each state, and that any information that moves the agent's beliefs is decision-relevant. Equivalently, the two conditions imply differentiability and strict convexity, respectively, of the agent's indirect utility function on the interior of the belief simplex. Appendix B.1 shows that regularity holds for a generic non-trivial problem with a finite action space when the agent's private information is imperfect but sufficiently rich.

Running Example. We use the following example throughout to illustrate our analysis. The agent is a doctor deciding whether a patient should undergo additional testing, $a = 1$, or not, $a = 0$, based on an interview and a preliminary test result that can be analyzed by AI (adviser).[11] The state is binary: $\omega = 1$ means that the patient is sick and $\omega = 0$ means that the patient is healthy. The doctor and AI receive conditionally independent signals, both inducing a full-support uniform distribution of posterior beliefs. We let $\theta \in [0, 1]$ denote the doctor's private belief and $\mu \in [0, 1]$ denote AI's belief. The doctor's payoff is $u(a, \omega) = \mathbf{1}_{a=\omega}$, i.e., the doctor would like to match the action to the state. It is easy to check that this decision problem is regular.

As a useful benchmark, consider the case in which the doctor has direct access to μ (equivalently, she fully trusts the adviser, i.e., $\alpha = 1$). By Bayes' rule, the posterior belief that the state is 1 after observing the realization (μ, θ) is $p(\mu, \theta) \triangleq \mu\theta/(\mu\theta + (1 - \mu)(1 - \theta))$. The doctor chooses $a = 1$ when $p(\mu, \theta) \geq 1/2$, or $\mu + \theta \geq 1$. Thus, under full trust, the doctor chooses additional testing whenever the combined evidence from her own assessment and the adviser's information favors the hypothesis that the patient is sick. Our goal is to understand how this decision rule changes when the adviser can only be trusted with probability $\alpha < 1$. □

[11] See, for example, Agarwal et al. (2025).

Remark 1. The assumption that s and θ are independent conditional on the state ω greatly simplifies the exposition and the interpretation of our results, but is effectively without loss of generality for our main characterization results. To see why, suppose that S is finite, and redefine the state space to be $\widetilde{\Omega} = \Omega \times S$. This makes the adviser's signal a deterministic function of the state, and hence trivially conditionally independent of θ. Identifying the adviser's information with his posterior over $\widetilde{\Omega}$, Theorems 1 and 2 hold *verbatim*.

3 Main Results

3.1 Trust Region Strategies

In what follows, it will be convenient to separate the dependence of the agent's strategy on the adviser's message and the agent's private information. To this end, we define a *private strategy* $\hat{\sigma}$ to be a function $\hat{\sigma} : \Theta \to \Delta(A)$ that specifies the agent's (possibly randomized) choice of action for each type. We will refer to the agent's belief about the state prior to updating based on her private type θ as the *interim belief*. If the agent has an interim belief μ and uses a private strategy $\hat{\sigma}$, her expected payoff is:

$$U(\hat{\sigma}, \mu) \triangleq \mathbb{E}_{\omega \sim \mu, \hat{\sigma}}[u(a, \omega, \theta)], \tag{3}$$

where the expectation is taken with respect to the conditional distribution of θ and the distribution of the agent's actions induced by $\hat{\sigma}$.[12] A private strategy $\hat{\sigma}$ is called *Bayes-optimal for belief* $\mu \in \Delta(\Omega)$ if it maximizes the agent's expected payoff when she holds an interim belief μ: $\hat{\sigma} \in \arg\max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$. A private strategy is called *Bayes-optimal* if it is Bayes-optimal for *some* belief. The agent's strategy can be viewed as a specification of a private strategy for each possible message received from the adviser, $\sigma \sim (\hat{\sigma}(m))_{m \in \Delta(\Omega)}$.

For a nonempty compact set $T \subseteq \Delta(\Omega)$ and $m \notin T$, let $\mathrm{Vis}(T, m)$ denote the part of the boundary of T visible from m. A point $p \in T$ is *visible* from m if the line segment from m to p does not intersect T before reaching p.[13]

[12] Formally, $\mathbb{E}_{\omega \sim \mu, \hat{\sigma}}[u(a, \omega, \theta)] = \sum_{\omega \in \Omega} \mu(\omega) \int_{\Theta} \int_{A} u(a, \omega, \theta) \hat{\sigma}(da|\theta) f(d\theta|\omega)$.

[13] Formally, $(1 - \lambda)m + \lambda p \notin T$ for every $\lambda \in [0, 1)$. Such a point belongs to the boundary of T.

Definition 1. $\sigma \sim (\hat{\sigma}(m))_{m\in\Delta(\Omega)}$ is a trust region strategy (TRS) if there exist a nonempty compact set $T \subseteq \Delta(\Omega)$ and a function $P : \Delta(\Omega) \to T$ such that

1. if $m \in T$, $\hat{\sigma}(m)$ is Bayes-optimal for m,

2. if $m \notin T$, $\hat{\sigma}(m)$ is Bayes-optimal for $P(m)$, where $P(m) \in \text{Vis}(T, m)$.

Intuitively, under a TRS, the agent treats messages m reported within the trust region T "at face value," i.e., she takes an optimal action treating m as her correct interim belief about the state. If a message m does not belong to the trust region T, the agent instead *projects* m onto the trust region, acting as if her interim belief were $P(m) \in T$ on the boundary of the trust region visible from m. Figure 3 in Section 5.2 provides a graphical illustration.

Running Example. In our running example, the space of beliefs is an interval: $\Delta(\Omega) = [0, 1]$. For illustration, suppose that the trust region is $T = [0.2, 0.8]$. Then, the corresponding trust region strategy takes the AI's reported belief m at face value if $m \in [0.2, 0.8]$ but otherwise replaces it with the nearest endpoint of the trust region, which is the only point on the visible part of the boundary. The corresponding decision rule is to choose additional testing, $a = 1$, whenever $\theta + \max\{0.2, \min\{m, 0.8\}\} \geq 1$. □

3.2 Optimality of Trust Region Strategies

We call two strategies of the agent *equivalent* if, together with some corresponding adviser's adversarial strategies, they induce the same joint distribution over states, types, messages, and actions. The importance of TRSs stems from the following key result.

Theorem 1 (Trust Region Solution). *An optimal strategy σ^* exists. Any optimal strategy is equivalent to a trust region strategy with a connected trust region T.*

Proof. See Appendix A.1 and A.2. □

Theorem 1 provides a sharp characterization of robustly optimal behavior under misalignment risk: Any optimal strategy can be interpreted as a TRS for some connected trust region T. Messages in the trust region are taken at face value while remaining messages are projected to their "safe interpretations" on the boundary of the trust region.

We first briefly sketch the intuition for Theorem 1. The proof is based on two key observations. First, the agent cannot benefit from using private strategies that are not Bayes-optimal for any interim belief. Indeed, the state-contingent expected payoffs induced by any such strategy are weakly dominated, state by state, by those of some Bayes-optimal private strategy; thus, replacing the former with the latter cannot hurt the agent, whether the adviser is aligned or misaligned. The trust region corresponds to interim beliefs that justify the set of Bayes-optimal strategies that the agent uses. Second, once that set is fixed, the matching of messages to interim beliefs can be constructed under the hypothetical assumption that the adviser is always aligned; this is because the misaligned adviser can achieve the same payoff for any such matching by simply adjusting which message to send to induce the target private strategy. Thus, each message is optimally mapped to a belief in the trust region that is "as close as possible," i.e., in the trust region if the message belongs to it and on the visible part of its boundary otherwise. As we show in Section 3.3, this intuition can be formalized using an appropriate notion of distance, which leads to a closed-form characterization of an optimal projection P.

Theorem 1 is silent about the optimal choice of the trust region itself. Two extreme cases are immediate. If the adviser is always aligned, a TRS with the trust region equal to the entire belief space is trivially optimal. If the adviser is always misaligned, an optimal TRS can be chosen with a trust region equal to the prior belief—the agent ignores the message of the adviser. In Section 3.5, we derive conditions under which the trust region is guaranteed to be non-trivial, so that the agent strictly benefits from the presence of the adviser.

In general, the shape of the optimal trust region can be quite complex, e.g., it need not be convex (see Figure 3 in Section 5.2). A trade-off is created by two opposing forces: expanding the set of available private strategies weakly increases the best payoff available at each truthful message but weakly decreases the payoff guaranteed against a misaligned adviser. Convexifying the trust region by adding beliefs between two trusted messages may lead some types of the misaligned adviser (especially those with very distant true beliefs) to use those newly trusted intermediate messages; the resulting losses may outweigh the gains from interpreting (nearby) truthful messages by the aligned adviser more accurately.

That said, Theorem 1 shows that the trust region may be chosen to be connected.[14] Moreover, the trust region can be chosen to have no "holes." Formally, we say that a set $A \subseteq \Delta(\Omega)$ is *non-hollow* if it contains all points x with the property that every line within the simplex going through x intersects A on both sides of x. Being non-hollow is implied by convexity but not by connectedness. An example of a connected but hollow set is a sphere; the corresponding ball is non-hollow.

Corollary 1. *Theorem 1 remains true with the additional requirement that T is non-hollow.*

Intuitively, the misaligned adviser's messages always induce interim beliefs on the boundary of the trust region; if a belief is "surrounded" from all directions by the trust region, it can be added to the trust region without modifying the trust region's boundary, and thus without decreasing the agent's worst-case payoff.

In general, the trust region may not be unique. So far, we emphasized that it can be taken to be a relatively large set. However, when the support M of the adviser's beliefs is finite, it is also possible to construct an optimal discrete trust region T with $|T| \leq |M|$. Intuitively, at most one belief in the trust region is needed for every possible belief of the aligned adviser. In such cases, a connected trust region can still be constructed but most beliefs in the trust region are never reported by the adviser. Uniqueness of the trust region can sometimes be established if the adviser's beliefs have full support (see Section 4).

Running Example. Theorem 1 implies that the doctor should follow a trust region strategy with a trust region that, by connectedness, can be taken to be an interval: $T = [\underline{\mu}, \bar{\mu}]$ for some $0 \leq \underline{\mu} \leq \bar{\mu} \leq 1$. It can be shown that the trust region $T = [0.2, 0.8]$ used previously for illustration is in fact optimal if (and only if) $\alpha = 55/59 \approx 0.93$; in Section 4, we derive a general formula for the cutoff beliefs $\underline{\mu}$ and $\bar{\mu}$. $\square$

[14] While the trust region cannot be convexified in the belief space, we instead establish convexity in dual coordinates, that is, in the space of state-contingent payoff profiles. For any fixed belief of the adviser, his adversarial payoff is *linear* in the state-contingent payoff profile induced by the agent's response. Thus, convexifying the set of state-contingent payoffs induced by the agent's private strategies maintains the payoff guarantee against misalignment. The enlarged set of payoffs corresponds to a connected set of beliefs at which one of its responses is Bayes-optimal.

3.3 Optimal boundary projection

The definition of a TRS requires the response to an untrusted message to be Bayes-optimal at some visible boundary belief, leaving many possible choices for the projection P. As we show next, there exists an economically interpretable characterization of the optimal projection.

With slight abuse of notation, let

$$U(\mu) \triangleq \max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$$

be the expected payoff to the agent when she uses a Bayes-optimal strategy at belief μ, and let $\text{BR}(\mu)$ denote the corresponding set of Bayes-optimal private strategies. For a TRS σ with trust region T, let $\text{Im}(\sigma)$ denote the set of (Bayes-optimal) private strategies it uses, and

$$d_\sigma(m, \mu) \triangleq U(m) - \sup_{\hat{\sigma} \in \text{BR}(\mu) \cap \text{Im}(\sigma)} U(\hat{\sigma}, m), \tag{4}$$

for all $\mu \in T$ and $m \in \Delta(\Omega)$. Intuitively, $d_\sigma(m, \mu)$ measures the agent's *regret* at belief m from restricting attention to Bayes-optimal strategies used by σ at belief $\mu \in T$, relative to using the private strategy optimal for belief m. Note that $d_\sigma(m, m) = 0$ for all $m \in T$ but $d_\sigma(m, \mu) \geq 0$ otherwise.

Corollary 2. *Every optimal strategy is equivalent to a TRS with a projection P satisfying, for every $m \notin T$,*

$$P(m) \in \arg\min_{\mu \in T} d_\sigma(m, \mu).$$

The corollary is a consequence of the proof of Theorem 1. Recall the intuition sketched in the preceding section: conditional on the set $\text{Im}(\sigma)$ of private strategies employed by the agent, the best she can do at a non-trusted message $m \notin T$ is to select a private strategy from $\text{Im}(\sigma)$ that maximizes her expected payoff under the hypothetical assumption that m is the true belief of the adviser. This is exactly what minimizing $d_\sigma(m, \mu)$ accomplishes: using m as her true interim belief, the agent looks for a trusted belief μ whose Bayes-optimal strategy comes closest to being optimal at m. In this sense, the agent maps m into the belief $\mu \in T$ that is a "closest safe interpretation" of the received message m.

The definition of d_σ is complicated by the fact that a TRS σ may use multiple Bayes-optimal strategies at some beliefs in the trust region. However, if the agent's problem is regular, every belief is associated with a unique Bayes-optimal payoff profile, and the characterization becomes sharper. In the regular case, the indirect payoff function U is differentiable on the interior of the belief simplex.[15] We write $\nabla U(\mu)$ for its gradient, which can be interpreted as the N-dimensional vector of state-contingent payoffs associated with a Bayes-optimal private strategy for μ.[16] In particular, $\nabla U(\mu) \cdot m$ is the expected payoff from the private strategy Bayes-optimal for belief μ when the agent holds belief m, where $\cdot$ denotes the standard dot product in $\mathbb{R}^N$. The extension is homogeneous of degree one, so Euler's identity gives $U(\mu) = \nabla U(\mu) \cdot \mu$, and thus

$$d_\sigma(m, \mu) = D_U(m, \mu) \triangleq U(m) - U(\mu) - \nabla U(\mu) \cdot (m - \mu), \tag{5}$$

where $D_U(m, \mu)$ is called the *Bregman distance* (Bregman (1967)), associated with function U, between beliefs m and μ. Thus, under the optimal TRS constructed above, messages outside the trust region are mapped into the belief in T that is closest in the Bregman distance:

$$P(m) \in \arg\min_{p \in T} D_U(m, p).$$

In other words, for regular decision problems, $P(m)$ can be taken to be the *Bregman projection* of m onto T.

Running Example. In our running example, $U(\mu) = 1 - \mu + \mu^2$ and $D_U(\mu, \mu') = (\mu - \mu')^2$; thus, Bregman projection coincides with the usual Euclidean projection and sends messages outside $[0.2, 0.8]$ to the nearest endpoint. □

[15] Since U is the supremum of functions linear in beliefs, and regularity ensures that all maximizers induce the same state-contingent payoff profile, differentiability of U in the interior of the belief simplex follows from the envelope theorem (see Milgrom and Segal (2002)).

[16] Formally, we extend U beyond the probability simplex by assuming that, for any non-negative measure μ, $U(\mu) = \mu(\Omega)\, U\left(\mu/\mu(\Omega)\right)$. At a boundary belief where an ordinary gradient is not defined, the expressions below use the unique Bayes-optimal payoff profile.

3.4 Robust Rationalizability

Our model assumes that the agent commits to a strategy at the outset of the game, not knowing the strategy adopted by the misaligned adviser. As we show next, the timing of moves does not matter for the value that the agent can achieve. This is because we can construct an optimal solution that is a saddle point of the zero-sum game between the agent and the misaligned adviser. For any strategy $\beta \in \mathcal{B}$ of the adversarial adviser, we let $\mathbb{P}_\beta(\cdot|m)$ denote the agent's interim belief induced by message m given the adviser's strategy.[17]

Definition 2 (Robustly Rationalizable Strategy). A strategy $\sigma^* \sim (\hat{\sigma}^*(m))_{m\in\Delta(\Omega)}$ is *robustly rationalizable* if there exists an adversarial strategy $\beta^* : M \to \Delta(M)$ of the misaligned adviser against σ^* such that for all $m \in M$, $\hat{\sigma}^*(m) \in \arg\max_{\hat{\sigma}'} U(\hat{\sigma}', \mathbb{P}_{\beta^*}(\cdot|m))$.

Suppose that the agent has a robustly rationalizable strategy σ^*. We argue that σ^* is optimal and, moreover, the agent cannot improve her worst-case payoff by observing the strategy chosen by the misaligned adviser. Suppose that the agent can observe the misaligned adviser's strategy β. Then, she can use Bayes' rule to compute her interim belief $\mathbb{P}_\beta(\cdot|m)$ following any realized message m, and choose a corresponding Bayes-optimal private strategy $\hat{\sigma}(m)$. This is clearly an optimal response to β, since it maximizes the agent's expected payoff conditional on every message. By definition, however, the robustly rationalizable strategy σ^* already prescribes a Bayes-optimal response after every message m against the worst-case reporting strategy β^*. This immediately implies that σ^* must be optimal: for any $\sigma \in \Sigma$,

$$V(\sigma) = \inf_{\beta\in\mathcal{B}} V(\beta, \sigma) \leq V(\beta^*, \sigma) \leq V(\beta^*, \sigma^*) = \min_{\beta\in\mathcal{B}} V(\beta, \sigma^*) = V(\sigma^*),$$

where the second inequality follows from the fact that σ^* responds optimally to β^*, and the second equality follows from the fact that β^* is adversarial against σ^*. In other words, the two strategies are mutual best responses in the zero-sum game between the agent and the misaligned adviser. It remains to establish that a robustly rationalizable strategy can always be found.

[17] A minimax reporting strategy can be chosen to use only messages in M; see Appendix A.1. The agent's robust guarantee continues to allow deviations to every message in $\Delta(\Omega)$.

Theorem 2 (Robust Rationalizability). *A robustly rationalizable strategy exists.*

Proof. See Appendix A.1 and A.3. □

Sketch of proof. The proof of Theorem 2 builds on Sion (1958)'s minimax theorem. However, the minimax theorem cannot be applied directly to our game when M is infinite because the players' payoffs do not, in general, satisfy the required continuity conditions. In a communication game, the meaning of messages is endogenous: when the game is infinite, two arbitrarily close messages may induce very different responses from the agent, and hence very different payoffs. Consequently, small changes in the adviser's strategy need not produce small changes in expected payoffs.

To overcome this difficulty, we use an alternative representation of the game in which the agent chooses a mapping from messages to state-contingent payoff profiles rather than private strategies, reducing the set of feasible choices at each message to a finite-dimensional set. Restricting the mapping to be continuous allows us to apply Sion's theorem and obtain an optimal reporting strategy for the misaligned adviser. We then show that any measurable mapping can be approximated arbitrarily well by continuous mappings in terms of expected payoff against the adversarial strategy. This establishes that the same adversarial strategy provides an upper bound on the agent's payoff even without the continuity restriction. A separate argument establishes the existence of an optimal strategy for the agent by exploiting compactness in the finite-dimensional space of state-contingent payoff profiles. Together, these arguments establish the existence of a saddle point.

When the sets M and Θ are finite, a more direct argument illuminates the economic intuition behind Theorem 2. For this argument, we restrict messages to M, suppose that $0 < \alpha < 1$, and assume that the Bayes-optimal payoff profile is unique at each induced interim belief. Suppose that the misaligned adviser moves before the agent, publicly committing to a reporting strategy β, so that the agent's best response is to use the Bayes-optimal private strategy at every realized message. Let β^* be the misaligned adviser's adversarial strategy given the agent's response. Under this timing of the game, the adviser cannot "deceive" the agent about how messages are generated: the agent directly observes the chosen information structure. Now consider the original timing, in which the agent commits to a decision rule

before the adviser chooses his reporting strategy. It would seem that the misaligned adviser could do strictly better in this case by manipulating the agent's beliefs covertly. For example, conditional on some true belief μ, the adviser might deviate from β^* and send a message that the agent mistakenly interprets as conveying a very different, potentially extreme belief, because she continues to update using β^*.

It turns out, however, that the misaligned adviser cannot benefit from a covert deviation relative to playing β^*. To see why, consider the first-order optimality condition for β^* in the game where the adviser moves first. Conditional on the adviser's belief μ, shifting a small probability mass from message m, sent with positive probability under β^*, to any alternative message $m' \in M$ cannot increase the misaligned adviser's payoff. Crucially, the envelope theorem implies that the first-order effect of this perturbation on the agent's optimized expected payoff can ignore the agent's adjustment of the optimal private strategy, and hence equals

$$(1-\alpha)\tau(\mu)\left[U(\hat{\sigma}^*(m'),\mu) - U(\hat{\sigma}^*(m),\mu)\right].$$

Since β^* is adversarial, the agent's payoff cannot decrease after any such perturbation, and hence, for every m sent with positive probability at μ and every alternative message m',

$$U(\hat{\sigma}^*(m'),\mu) \geq U(\hat{\sigma}^*(m),\mu).$$

But this is precisely the condition that rules out a profitable covert deviation: holding fixed the agent's response to each message, the misaligned adviser cannot further reduce her expected payoff by switching from the on-path m to any alternative m'. Although the agent's ability to adjust her beliefs after an overt deviation might seem to discipline the adviser, the envelope theorem shows that this adjustment has no first-order effect on her optimized payoff. Thus, the adversarial reporting strategy in the overt-information game already satisfies the incentive constraints of the covert-information game. Consequently, the same strategies form an equilibrium under the original timing.

Trust region equilibria. Let σ^* be robustly rationalizable. In light of Theorem 1, the agent's equilibrium strategy can be taken to be a TRS. Treating the problem as a zero-sum

game between the agent and the misaligned adviser, we will call (σ^*, β^*) a *trust region equilibrium* (TRE) if σ^* is a TRS that is robustly rationalizable against the adversarial strategy β^*.

Suppose that the agent's decision problem is regular. In a TRE, messages $m \in M$ in the trust region are taken at face value (thus, Bayes' rule implies that $\mathbb{P}_{\beta^*}(\cdot|m) = m$). Messages $m \in M$ outside of the trust region are reported by both types of the adviser with probabilities such that $\mathbb{P}_{\beta^*}(\cdot|m) = P(m)$, where $P(m)$ is the Bregman projection onto the trust region described in Section 3.3. Messages $m \notin M$ are sent with probability zero. In other words, the mappings from messages to beliefs induced by *(i)* Bayes' rule and *(ii)* Bregman projection onto the trust region *coincide* on the equilibrium path of a TRE. In Section 4, we use this structural property to characterize the trust region in a binary-state setting.

Running Example. We can derive a robustly rationalizable strategy explicitly in the running example. Let $\alpha = 55/59$, so that the trust region $T = [0.2, 0.8]$ is optimal for the doctor. Define β^* as follows: when the true belief $\mu \leq 1/2$, the misaligned adviser sends a random message $m \in [0.8, 1]$ with density $50(m - 0.8)$; when $\mu > 1/2$, he sends a random message $m \in [0, 0.2]$ with density $50(0.2 - m)$. Under this strategy, Bayes' rule implies that the doctor's interim belief after observing message m is m when $m \in (0.2, 0.8)$ (these messages are never sent by the misaligned adviser), 0.8 when $m \geq 0.8$, and 0.2 when $m \leq 0.2$. Thus, the doctor's optimal response to β^* is exactly her trust-region strategy. Conversely, β^* is a best response for the misaligned adviser, since this strategy always induces the most harmful belief within the trust region. Hence, the two strategies form a saddle point.

The example illustrates that worst-case misalignment need not take an implausible form. Here, the misaligned AI simply predicts the state opposite to that suggested by its information and reports this prediction with exaggerated, though not absolute, confidence. □

Implications. From a technical perspective, Theorem 2 provides a practical way of certifying the optimality of solutions in applications. It suffices to construct a saddle point of the zero-sum game between the agent and the misaligned adviser. Verifying the mutual best-response property is often easier than evaluating the agent's objective for every possible strategy. In particular, we use this method to construct the optimal trust region numerically in Section 5.

Finally, Theorem 2 implies that our problem is equivalent to a *constrained persuasion problem* for the misaligned adviser. When the misaligned adviser moves first (i.e., we are in the overt-information game), he is effectively choosing a Bayes-plausible distribution of the agent's interim beliefs subject to the constraint that truthful reporting occurs with probability at least α after each adviser's signal; the constraint reflects the presence of the aligned adviser. Thus, the misaligned adviser is effectively attempting to "jam" the signal sent by the aligned adviser. We exploit this perspective in the next section to derive thresholds on α below which no TRE can sustain informative communication.

3.5 Minimal Viable Alignment

In this section, we derive bounds on the alignment probability α above which the agent finds it worthwhile to consult the adviser. Formally, define the value of an adviser as

$$\Delta V \triangleq V^* - V_0,$$

where $V_0 \triangleq \sup_{\sigma \in \Sigma} \mathbb{E}_\sigma[u(a, \omega, \theta)]$ is the agent's optimal payoff in the absence of the adviser. Since the agent can always ignore the adviser's messages, this value is non-negative, $\Delta V \geq 0$. We ask when this value is strictly positive, $\Delta V > 0$ (cf. the value of information by Blackwell (1951)).

To answer this question, we derive a bound on α that is independent of the agent's problem. If α is small enough, the misaligned adviser can use a strategy β that "jams" the signal created by truthful reporting of the aligned adviser. In such a case, the distribution of interim beliefs $\mathbb{P}_\beta(\cdot|m)$ held by the agent is degenerate. An optimal strategy can then ignore the adviser and be represented by the singleton trust region $\{\mu_0\}$. However, if α is large enough, there exists no strategy for the misaligned adviser that makes the equilibrium message uninformative. In such cases, as long as information is useful to the agent (the agent's problem is regular), ΔV must be strictly positive.

Definition 3 (Minimal Viable Alignment). The minimal viable alignment $\text{MVA}(\tau)$ is the smallest upper bound on α for which there exists a strategy β of the misaligned adviser such that the induced interim belief satisfies $\mathbb{P}_\beta(\cdot|m) = \mu_0$ for every $m \in M$.

MVA depends on the adviser's information $\tau \in \Delta(\Delta(\Omega))$. Define the rank of this information as

$$R(\tau) \triangleq \dim \operatorname{span}(M). \tag{6}$$

For finite support $M = \{\mu_1, \ldots, \mu_K\}$, this is the rank of the matrix whose columns are the adviser's posteriors. Intuitively, $R(\tau)$ is related to the richness of the adviser's information: the adviser's beliefs are located in an $(R(\tau)-1)$-dimensional subspace of the $(N-1)$-dimensional belief simplex $\Delta(\Omega)$. For any τ, $R(\tau) \leq N$; with finite support, $R(\tau) \leq \min\{|M|, N\}$. The rank weakly decreases when the adviser's information is garbled. In what follows, we assume that the adviser has some information, so $R(\tau) \geq 2$.

Theorem 3 (Minimal Viable Alignment). *The agent strictly benefits from the presence of the adviser, $\Delta V > 0$, in some decision problem (equivalently, in any regular decision problem) if and only if $\alpha > \mathrm{MVA}(\tau)$. For any τ, $\mathrm{MVA}(\tau) \in [1/R(\tau), 1/2] \subseteq [1/N, 1/2]$. If $R(\tau) = |M| < \infty$, then $\mathrm{MVA}(\tau) = 1/R(\tau)$. Moreover, for any $\alpha \in [1/N, 1/2]$, there exists τ such that $\mathrm{MVA}(\tau) = \alpha$.*

Proof. See Appendix A.4. □

When M is finite, the proof of Theorem 3 shows that $\mathrm{MVA}(\tau)$ can be computed as the solution to a finite-dimensional linear program. We establish bounds on this solution and then construct adviser information structures attaining every value in the universal range $[1/N, 1/2]$.

The role of rank is clearest when the K adviser posteriors are linearly independent. To make messages uninformative about the state, the misaligned adviser must then make the overall distribution of messages identical across the adviser's signals. If p_j is the probability of message j conditional on any signal, truthful reporting requires $p_j \geq \alpha$ for every j, so $K\alpha \leq 1$. Conversely, whenever this inequality holds, the misaligned adviser can fill in the remaining probabilities so that every signal induces the uniform distribution over messages. Hence the threshold is $1/K = 1/R(\tau)$.

By Theorem 3, if the alignment α exceeds $1/2$ and information is strictly useful everywhere, then the agent always benefits from the presence of the adviser. Conversely, if the state

is binary and $\alpha < 1/2$, the agent cannot benefit from the adviser. In higher-dimensional problems, the adviser can be valuable at much lower alignment levels. In particular, if $R(\tau) = K = N$, then it suffices that $\alpha > 1/N$. Thus, when the state space is very rich, even a small amount of trust is enough for the agent to benefit from the advice of a misaligned adviser.

In regular decision problems, the MVA is the alignment threshold above which the value of advice ΔV becomes strictly positive. More generally, ΔV is convex in α and therefore strictly increasing whenever positive. Convexity follows because, for each fixed decision rule σ, the agent's objective is affine in α—a weighted sum of the payoff under truthful reporting and under worst-case misalignment—and optimizing over decision rules corresponds to taking the upper envelope of these functions. An economic implication is that under the robust optimality criterion, investment in alignment has non-decreasing marginal returns. Intuitively, as the agent's reliance on AI's feedback increases, the marginal gain from being in the aligned case gets larger.[18]

Running Example. Because the state is binary and the decision problem is regular in our running example, the doctor should not use AI assistance if $\alpha \leq 1/2$: the trust region is a singleton containing the prior, $T = \{\mu_0\}$. For any $\alpha > 1/2$, the trust region is a non-degenerate interval. The doctor's value of an adviser ΔV is zero for $\alpha \leq 1/2$, strictly positive and convex for $\alpha \in (1/2, 1]$, with a right derivative of 0 at $\alpha = 1/2$ and a left derivative of $7/12$ at $\alpha = 1$. $\square$

4 Binary State

Consider the case of a binary state, $\Omega = \{0, 1\}$. We can intuitively think of the state as capturing whether a given statement is false or true—as in our running example. The belief is effectively one-dimensional: with slight abuse of notation, let $\mu \in [0, 1]$ denote the probability of state $\omega = 1$. We assume that the adviser's posterior is distributed over $M = [0, 1]$ with a strictly positive, continuous probability density $\tau(\mu)$.

[18] Of course, the implications of this observation for the profitability of investment in alignment depend on the marginal cost function which is likely also convex in α.

Throughout this section, we assume that the agent's decision problem is regular. Regularity implies that the indirect payoff function $U(\mu)$ is differentiable and strictly convex on $(0,1)$. For expositional clarity, we further assume that $U(\mu)$ is twice continuously differentiable on $[0,1]$ with $U''(\mu) > 0$ for all $\mu \in (0,1)$. We write $\hat{\sigma}_0(\mu)$ for a Bayes-optimal private strategy.

Since any connected one-dimensional compact set is a closed interval, a straightforward corollary of Theorem 1 is:

Corollary 3. *If $|\Omega| = 2$, any optimal strategy σ^* is equivalent to a trust region strategy with $T = [\underline{\mu}, \overline{\mu}]$. If $m \in [\underline{\mu}, \overline{\mu}]$, $\hat{\sigma}(m) = \hat{\sigma}_0(m)$; if $m < \underline{\mu}$, $\hat{\sigma}(m) = \hat{\sigma}_0(\underline{\mu})$; if $m > \overline{\mu}$, $\hat{\sigma}(m) = \hat{\sigma}_0(\overline{\mu})$.*

If $\underline{\mu} = \overline{\mu}$, the agent effectively ignores the adviser, implying $\underline{\mu} = \overline{\mu} = \mu_0$. If $\underline{\mu} < \overline{\mu}$, the agent plays according to the Bayes-optimal strategy $\hat{\sigma}_0(\underline{\mu})$ if $m \leq \underline{\mu}$, and according to the Bayes-optimal strategy $\hat{\sigma}_0(\overline{\mu})$ if $m \geq \overline{\mu}$.

When the trust region is an interval, the misaligned adviser optimally induces one of the two endpoint beliefs, depending on his own realized posterior (as in the running example); thus the worst-case misaligned behavior admits a simple threshold characterization:

Lemma 1. *When the agent commits to a TRS with the trust region $T = [\underline{\mu}, \overline{\mu}]$, where $\underline{\mu} < \overline{\mu}$, the misaligned adviser with belief $\mu \in [0,1]$ finds it optimal to send any message $m \geq \overline{\mu}$ if $\mu \leq b(\underline{\mu}, \overline{\mu})$ and any message $m \leq \underline{\mu}$ if $\mu \geq b(\underline{\mu}, \overline{\mu})$, where*

$$b(\underline{\mu}, \overline{\mu}) = \frac{\int_{\underline{\mu}}^{\overline{\mu}} xU''(x)dx}{\int_{\underline{\mu}}^{\overline{\mu}} U''(x)dx}. \tag{7}$$

Proof. See Appendix A.5. □

Lemma 1 states that the misaligned adviser with high enough beliefs μ will induce the private strategy Bayes-optimal at the lowest belief in the trust region, $\underline{\mu}$, by reporting some message m lower than $\underline{\mu}$; similarly, the misaligned adviser with low enough beliefs μ will induce the private strategy Bayes-optimal at the highest belief in the trust region, $\overline{\mu}$, by reporting some message m higher than $\overline{\mu}$. The threshold belief is given by the conditional expectation of a random variable whose distribution is determined by the curvature of the indirect utility function: $b(\underline{\mu}, \overline{\mu}) = \mathbb{E}[\nu | \nu \in [\underline{\mu}, \overline{\mu}]]$, where ν is distributed with full support

over $[0,1]$ according to probability density $U''(\cdot)/\int_0^1 U''(x)dx$. Since U'' measures the local value of information, the cutoff gives greater weight to beliefs at which information is more valuable to the agent.

To characterize the trust region's boundaries, we will use the observation from Theorem 2 that it is sufficient to construct mutual best responses for the agent and the misaligned adviser. Lemma 1 characterizes the best response of the misaligned adviser. A best response of the agent must use the Bayes-optimal strategies at each interim belief induced by the adviser's strategy. A necessary condition is that the average interim belief induced by messages $m \leq \underline{\mu}$ is exactly $\underline{\mu}$, and the average interim belief induced by messages $m \geq \overline{\mu}$ is exactly $\overline{\mu}$:[19]

$$\frac{\alpha \int_0^{\underline{\mu}} \mu\tau(\mu)d\mu + (1-\alpha)\int_{b(\underline{\mu},\overline{\mu})}^1 \mu\tau(\mu)d\mu}{\alpha \int_0^{\underline{\mu}} \tau(\mu)d\mu + (1-\alpha)\int_{b(\underline{\mu},\overline{\mu})}^1 \tau(\mu)d\mu} = \underline{\mu}, \tag{8}$$

$$\frac{\alpha \int_{\overline{\mu}}^1 \mu\tau(\mu)d\mu + (1-\alpha)\int_0^{b(\underline{\mu},\overline{\mu})} \mu\tau(\mu)d\mu}{\alpha \int_{\overline{\mu}}^1 \tau(\mu)d\mu + (1-\alpha)\int_0^{b(\underline{\mu},\overline{\mu})} \tau(\mu)d\mu} = \overline{\mu}. \tag{9}$$

As it turns out, these conditions are also sufficient for a TRE:

Proposition 1 (Binary-State Characterization). *An optimal strategy exists; it is unique and robustly rationalizable. Its trust region, $T = [\underline{\mu}, \overline{\mu}]$, is equal to the prior belief $\{\mu_0\}$ when $\alpha \leq 1/2$; when $\alpha \in (1/2, 1)$, it is defined by the unique solution to the system (8)-(9) that satisfies $\underline{\mu} \leq \mu_0 \leq \overline{\mu}$; when $\alpha = 1$, $T = [0,1]$.*

Proof. See Appendix A.6. □

Note that while the structure of the trust region characterized by Proposition 1 is simple, the underlying strategy of the misaligned adviser may be quite complex in a TRE. By Lemma 1, the misaligned adviser with belief $\mu \geq b(\underline{\mu}, \overline{\mu})$ is indifferent between sending all messages $m \leq \underline{\mu}$ since they all result in the same Bayes-optimal strategy $\hat{\sigma}_0(\underline{\mu})$. In a commitment solution, the misaligned adviser can send any of these messages; for example, he can always send $m = \underline{\mu}$. But in a TRE, the strategy β^* of the misaligned adviser must be such that *every* message $m \leq \underline{\mu}$ induces the interim belief $\underline{\mu}$ via Bayes' rule. Since all messages $m \leq \underline{\mu}$

[19] One way to see that is to use our observation that in a TRE, the mapping from messages to belief defined by Bayes' rule must agree with the mapping defined by minimizing the Bregman distance to the trust region.

are sent on the equilibrium path (the aligned adviser simply reports his belief truthfully), β^* must rely on the misaligned adviser's indifference to put just enough probability mass on each of these messages to induce $\underline{\mu}$.

Proposition 1 fully characterizes the optimal trust region. A natural next question is how the trust region depends on the problem's parameters. We offer two comparative statics results, one related to the size of the trust region, and one related to its location.

First, higher alignment results in more trust:

Proposition 2 (Change in Alignment). *When $\alpha \geq 1/2$, $\underline{\mu}(\alpha)$ is strictly and continuously decreasing in α and $\overline{\mu}(\alpha)$ is strictly and continuously increasing in α.*

Proof. See Appendix A.7. □

Proposition 2 shows that the trust region gradually and monotonically expands from the prior belief, at $\alpha \leq 1/2$, to the entire belief simplex. For any $\alpha < 1$, the trust region excludes the most extreme beliefs. Intuitively, the aligned adviser is unlikely to hold such beliefs, whereas the misaligned adviser would exploit their inclusion in the trust region to induce the corresponding extreme responses (by Lemma 1).

Second, we show that the trust region tends to include beliefs at which the decision problem of the agent is less "information-sensitive." In other words, the agent will avoid expanding the trust region to beliefs where small changes in information lead to large changes in the optimal action. We formalize this notion via the indirect utility function $U(\mu)$, noting that its curvature reflects the sensitivity of the agent's optimal private strategy to her interim beliefs.

Definition 4 (Information Sensitivity). We say that the indirect utility function $U_1(\mu)$ is *less information-sensitive at higher beliefs* than the indirect utility function $U_2(\mu)$ if

$$\frac{U_1''(\mu)}{U_2''(\mu)} \text{ is decreasing in } \mu.$$

The definition compares the local value of information in the two problems: the curvature of U_1 relative to that of U_2 is smaller at higher beliefs μ. Intuitively, under U_1, the decision

of the agent is less sensitive to new information at higher beliefs. It turns out that in this case the trust region will be skewed towards higher beliefs.

Proposition 3 (Change in Information Sensitivity). *Suppose that $U_1(\mu)$ is less information-sensitive at higher beliefs than $U_2(\mu)$. Then, the trust region T_1 corresponding to U_1 is higher in the strong set order than the trust region T_2 corresponding to U_2.*

Proof. See Appendix A.8. □

Proposition 3 shows that the trust region skews towards beliefs at which the agent's optimal action is less sensitive to new information.

Running Example. Applying Proposition 1, we obtain a symmetric trust region $T_\alpha = [\underline{\mu}(\alpha), 1 - \underline{\mu}(\alpha)]$, where

$$\underline{\mu}(\alpha) = \begin{cases} \frac{1}{2} & \alpha \leq \frac{1}{2}, \\ \frac{\sqrt{(1-\alpha)(1+2\alpha)}-(1-\alpha)}{2\alpha} & \alpha > \frac{1}{2}. \end{cases}$$

Figure 1 depicts the optimal trust region as a function of α, while Figure 2 illustrates the resulting decision rule for the doctor as a function of the realized signals. In line with Proposition 1 and Proposition 2, the trust region is equal to the prior belief when the alignment probability is below 1/2. In that case, the doctor should not use AI and instead rely exclusively on her own signal, as shown in the left panel of Figure 2. When the alignment probability is above 1/2, the doctor trusts moderate AI messages. Extreme messages, namely those with $m < \underline{\mu}(\alpha)$ or $m > 1 - \underline{\mu}(\alpha)$, are projected to the endpoints of the trust region. The optimal decision rule is therefore sensitive to AI's recommendations only in the intermediate range of beliefs. In particular, the AI's signal alone is never sufficient to induce testing without corroboration from the doctor's information, as shown in the middle panel of Figure 2. Finally, as α approaches 1, the trust region converges to the entire simplex (Figure 1), and the optimal decision rule converges to the first-best decision rule, as shown in the right panel of Figure 2.

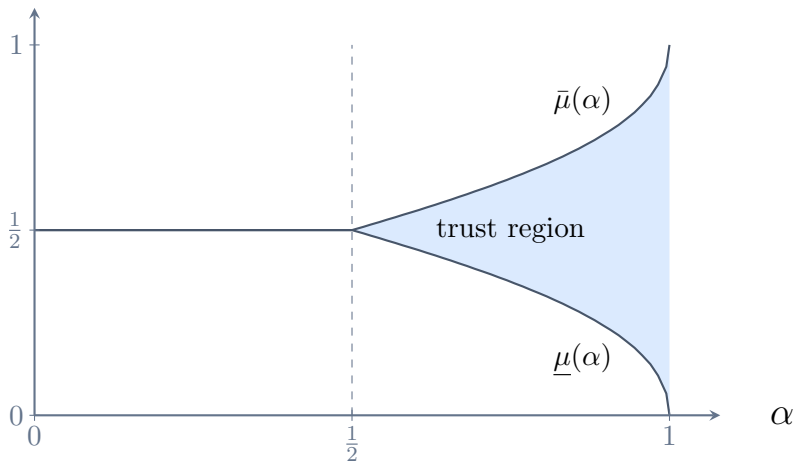


Figure 1: The optimal trust region as a function of alignment probability α.

Asymmetric stakes. Suppose that the doctor's payoff is modified to

$u(a,\omega)$	$\omega=0$	$\omega=1$
$a=0$	1	0
$a=1$	0	r

.

We assume that $r > 1$, i.e., payoffs are more sensitive to taking the correct action when the patient is sick. The system of equations (7)-(9) no longer admits a closed-form solution. However, as r increases, the doctor's indirect payoff $U(\mu) = r\mu + (1-\mu)^2/(1+(r-1)\mu)$ becomes less information-sensitive at higher beliefs. Intuitively, when $r > 1$, the objective function makes the doctor effectively place more weight on state 1, in which the patient is sick. As a result, she reacts less strongly to new information when she believes that state 1 is likely. By Proposition 3, increasing r shifts the optimal trust region upward. The doctor therefore trusts higher messages more than before and becomes more skeptical of low messages.

Takeaways. The example resonates with how imaging AI is often deployed in practice. Most real-world systems are used as second readers or prioritization aids rather than autonomous

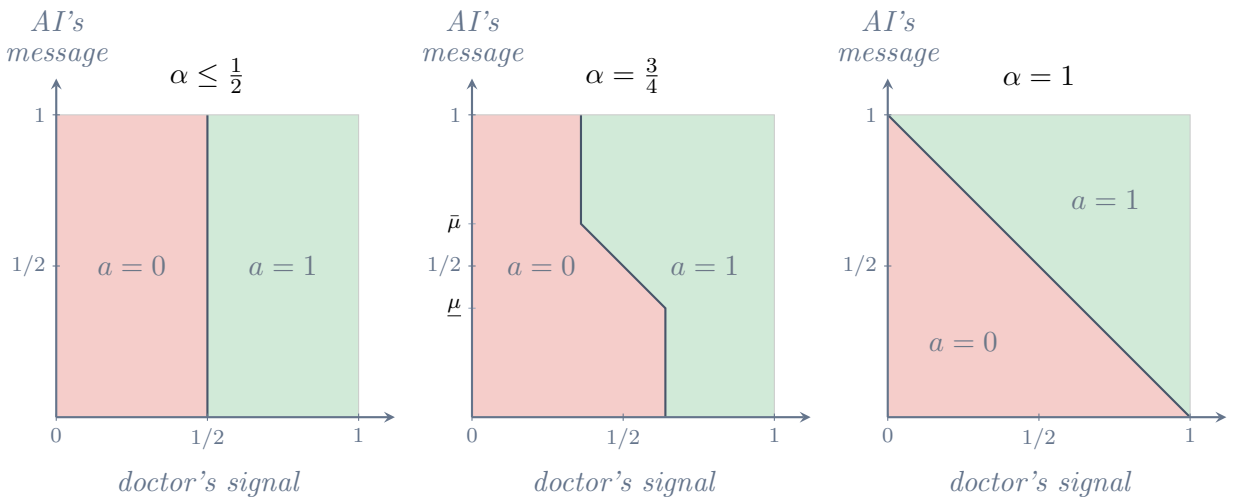


Figure 2: Optimal decision rule in three cases: $\alpha \leq 1/2$, $\alpha = 3/4$, and $\alpha = 1$.

decision-makers.[20] The same caution is consistent with growing evidence on automation bias: when clinicians are shown incorrect AI suggestions, their decisions can deteriorate.[21]

Our analysis offers a normative rationale for such limits on the influence of AI advice. Under the trust-region protocol, intermediate probabilities should be used as reported, whereas extremely low or extremely high probabilities should be replaced by the corresponding boundary values of the trusted range. Thus even a near-certain AI prediction leaves room for contrary evidence from the doctor's own assessment. □

5 Multiple States

In this section, we consider the general case $|\Omega| \geq 2$. Except for cases exhibiting special structure (see Appendix B), an analytical characterization of the optimal trust region is no longer available. We therefore develop a numerical procedure that can be used for any finite M, A and Θ, and then apply it to an extension of the medical triage example.

[20] A recent systematic review of real-world AI deployment in medical imaging finds that AI most commonly serves as a secondary reader or a triage tool rather than as a fully autonomous reader; see Wenderott et al. (2024). For a regulatory example, the FDA-cleared DrAid chest X-ray system is indicated as a triage and prioritization aid and is not intended for stand-alone clinical decision-making; see FDA 510(k) summary K221241.

[21] See Dratsch et al. (2023). An RSNA summary of the study reports that incorrect AI suggestions reduced reader accuracy even for experienced radiologists.

5.1 Finding the trust region numerically

Optimizing over trust regions directly is not computationally tractable, as the problem is non-linear and high-dimensional. Instead, we exploit our minimax theorem (Theorem 2).[22] Equilibrium strategies can be found by solving a pair of linear programs—the primal, corresponding to finding the optimal strategy for the agent, and the dual, corresponding to finding the adversarial strategy for the adviser. The trust region is then pinned down by these strategies. Additionally, we describe a simple procedure for enlarging the initial trust region to a connected one.[23]

Throughout, we focus on the non-degenerate case $\alpha \in (0,1)$. Suppose that $\Theta = \{\theta_1, \ldots, \theta_K\}$, $A = \{a_1, \ldots, a_J\}$, and $M = \{\mu_1, \ldots, \mu_I\}$. Let $\tau_i \equiv \tau(\mu_i) > 0$ be the probability that the adviser holds belief μ_i. For each posterior μ_i, type θ, and action a, define

$$u_{i\theta a} \triangleq \sum_{\omega \in \Omega} \mu_i(\omega) f(\theta|\omega) u(a, \omega, \theta)$$

as the contribution to expected utility from type θ choosing action a, conditional on the adviser's posterior being μ_i. The agent's problem can then be written as the following linear program:

$$\begin{aligned} V^* = \max_{\sigma, \underline{U}} \quad & \sum_i \tau_i \left[\alpha \sum_{\theta,a} u_{i\theta a} \sigma_{i\theta a} + (1-\alpha) \underline{U}_i \right] \\ \text{s.t.} \quad & \sum_a \sigma_{j\theta a} = 1 \ (\forall j, \theta), \quad \sigma_{j\theta a} \geq 0 \ (\forall j, \theta, a), \quad \underline{U}_i \leq \sum_{\theta,a} u_{i\theta a} \sigma_{j\theta a} \ (\forall i, j). \end{aligned} \tag{P}$$

Here, the variable $\sigma_{j\theta a}$ captures the agent's strategy: it is the probability that the agent chooses action a after observing message μ_j and type θ. Thus, if the adviser's posterior is μ_i but the message is μ_j, the agent's expected payoff is $\sum_{\theta,a} u_{i\theta a} \sigma_{j\theta a}$. The first two constraints ensure that the strategy is well defined. The auxiliary variable $\underline{U}_i$ can be interpreted as

[22] This use of the minimax theorem to simplify a strategic communication problem is related in spirit to Lipnowski and Ravid (2020), who use a max-min characterization of equilibrium payoffs to transform a cheap-talk problem into an optimization problem.

[23] The trust region will always be a discrete set in a numerical solution, but the procedure yields a trust region that can be as close to being connected as the numerical approximation allows.

the payoff guaranteed to the agent conditional on the adviser's posterior being μ_i. The last constraint ensures that, at an optimum,

$$\underline{U}_i = \min_j \sum_{\theta,a} u_{i\theta a}\sigma_{j\theta a},$$

so that $\underline{U}_i$ is precisely the payoff obtained against the worst-case message of a misaligned adviser with posterior μ_i.

The dual of (P) can be written as

$$\begin{aligned} V^* = \min_{\beta,\overline{U}} \quad & \sum_{j,\theta} \overline{U}_{j\theta} \qquad\qquad\qquad\

message μ_j is

$$\widehat{\mu}_j \triangleq \frac{\sum_i \mu_i \tau_i \left[\alpha \mathbf{1}_{i=j} + (1-\alpha)\beta^*_{ij}\right]}{\sum_i \tau_i \left[\alpha \mathbf{1}_{i=j} + (1-\alpha)\beta^*_{ij}\right]}. \tag{10}$$

Complementary slackness implies that the private strategy prescribed by σ^* following μ_j is Bayes-optimal at $\widehat{\mu}_j$. Then,

$$T^{\text{imp}} \triangleq \{\widehat{\mu}_1, \ldots, \widehat{\mu}_I\}$$

is the set of beliefs implemented on the equilibrium path. It is easy to show that there exists an optimal TRS for the agent with the trust region T^{imp}.

Next, we explain how to recover a *connected* trust region by enlarging the set T^{imp}. Let W be the set of all state-contingent payoff profiles that are feasible for the agent, i.e., induced by some private strategy (as defined in Appendix A.1). The optimal guarantees $\underline{U}^*_i$ determine which interim beliefs can be safely trusted, in the sense that including them in the trust region does not allow any type of the misaligned adviser to decrease the agent's payoff below the corresponding guarantee. Define

$$\overline{T} \triangleq \left\{\mu \in \Delta(\Omega) : \ \exists\, w \in W \text{ such that } \mu \cdot w = \max_{\widetilde{w} \in W} \mu \cdot \widetilde{w} \text{ and } \mu_i \cdot w \geq \underline{U}^*_i, \forall i \right\}.$$

Intuitively, we add a belief μ to $\overline{T}$ if there exists a private strategy for the agent—associated with a feasible payoff profile $w \in W$—such that the strategy is Bayes-optimal for μ (the first condition) and cannot be exploited by a misaligned adviser with belief μ_i to bring the agent's payoff below the guarantee $\underline{U}^*_i$ (the second condition). The set $\overline{T}$ contains T^{imp}: the private strategy $\hat{\sigma}_j$ defined by $\hat{\sigma}_j(\theta)(a) \triangleq \sigma^*_{j,\theta,a}$, for each θ and a, is Bayes-optimal for $\widehat{\mu}_j$, while its associated payoff profile satisfies the second condition by primal feasibility of σ^*. By construction, $\overline{T}$ is the largest, in the sense of inclusion, set of beliefs containing T^{imp} that preserves the agent's payoff conditional on misalignment. Selecting a payoff-maximizing response from this enlarged set at every message weakly increases the agent's payoff conditional on alignment, so the resulting strategy remains optimal. Finally, $\overline{T}$ is connected by the argument used in the proof of Theorem 1.

The construction translates directly into a numerical procedure for approximating the

trust region on an arbitrarily fine grid of beliefs in $\Delta(\Omega)$. The feasible set W is a polytope, and the subset of payoff profiles satisfying the constraints $\mu_i \cdot w \geq \underline{U}_i^*$, $\forall i$, is itself a polytope represented by finitely many linear inequalities. Thus, after solving the primal and dual programs, one can construct this finite system of inequalities once and then test each candidate belief μ independently: μ is included in the numerical approximation to $\overline{T}$ if some Bayes-optimal payoff profile at μ satisfies all of the inequalities.

5.2 Medical Triage Example Revisited

We illustrate the numerical procedure by solving an extended version of the running medical triage example. In addition to the two states, which we now label "Healthy" and "Sick," there is a third state called "Other (condition)." The doctor's payoffs from action $a = 1$ (testing) are 1 in "Sick," κ in "Other," and 0 in "Healthy." No testing gives a payoff of 1 if the patient is healthy and 0 otherwise. The new parameter κ could be positive or negative, reflecting either an alternative diagnosis with lower stakes ($0 < \kappa < 1$) or a condition for which testing is harmful ($\kappa < 0$). We again assume that the doctor's and AI's private signals are conditionally independent and induce a uniform distribution over the simplex in the baseline case.[24] We also look at the case in which the doctor's realized signal is equal to the baseline signal with probability γ and a signal uncorrelated with the state, with the same marginal distribution, with probability $1 - \gamma$. Then, $\gamma = 1$ corresponds to the doctor being ex-ante equally well-informed as AI, while lower γ allows us to capture cases in which AI is more informed.

Figure 3 depicts an optimal trust region T for the baseline scenario with $\alpha = 0.9$, $\kappa = 0.5$, and $\gamma = 1$. The left panel illustrates the function $P(m)$, which maps each received message m outside the trust region to a trusted belief on its boundary. Recall from Section 3.3 that, under regularity, this mapping can be chosen as the Bregman projection of m onto T, with Bregman distance induced by the indirect utility function U of the agent. The sets of messages admitting a given Bregman projection are convex: if a belief $\mu \in T$ is the Bregman projection of both m and m', then μ is also the Bregman projection of every convex

[24] For the numerical procedure, we work with a fine discrete approximation of the uniform distribution.

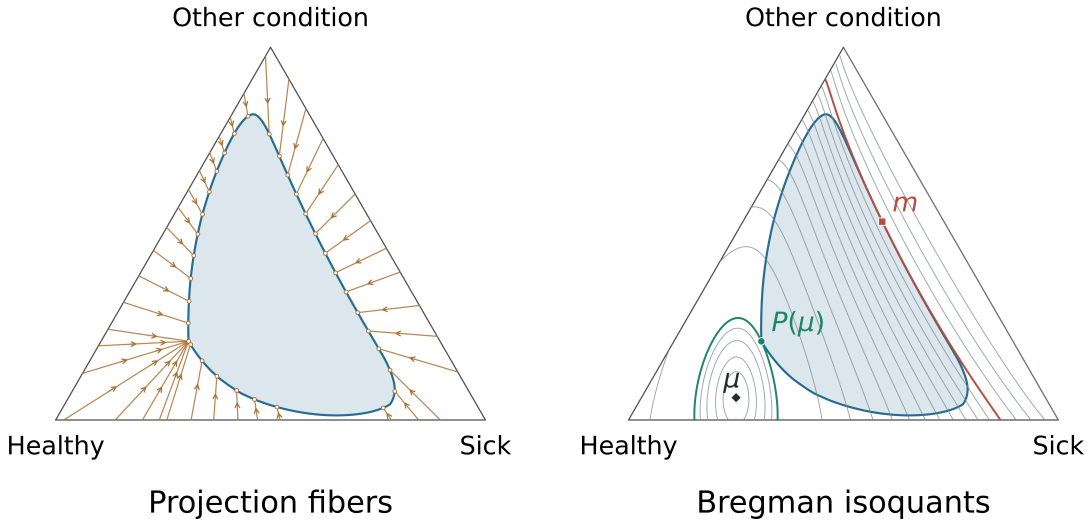


Figure 3: Trust region in the baseline case $\alpha = 0.9$, $\kappa = 0.5$, and $\gamma = 1$.

combination $\lambda m + (1 - \lambda)m'$, $\lambda \in [0, 1]$, by equation (5).

The right panel of Figure 3 depicts the isoquants (level sets) of the Bregman distance from posterior belief μ. Suppose that the adviser holds posterior belief μ. If the adviser is aligned, he will report μ, which the agent takes at face value if it lies in T and otherwise maps to $P(\mu)$. If the adviser is misaligned, a simple calculation shows that he will instead report a message inducing a belief within the trust region that *maximizes* the Bregman distance to μ, point m in Figure 3. Thus, the geometry of the trust region is governed by the shape of the level sets of the Bregman distance induced by U.

Unlike in the simplest version of our running example, Bregman distance is not a (square of a) metric in general, as it may not satisfy the triangle inequality or symmetry. A maximizer of the Bregman distance need not be an extreme point of T, but can be chosen on its boundary, because the distance weakly increases along rays from the true belief. Consequently, the geometry of T may be quite complex; for example, in Figure 3, the optimal T is not convex. Adding some belief in the convex hull of T would benefit the agent when the adviser is aligned, but would allow the misaligned adviser with posteriors such as μ to induce beliefs that are further away from the truth. Such trade-offs make an analytical characterization of the optimal trust region infeasible in general. In Appendix B.2.2, we solve a special multi-dimensional example in which the isoquants of the Bregman distance are spheres.

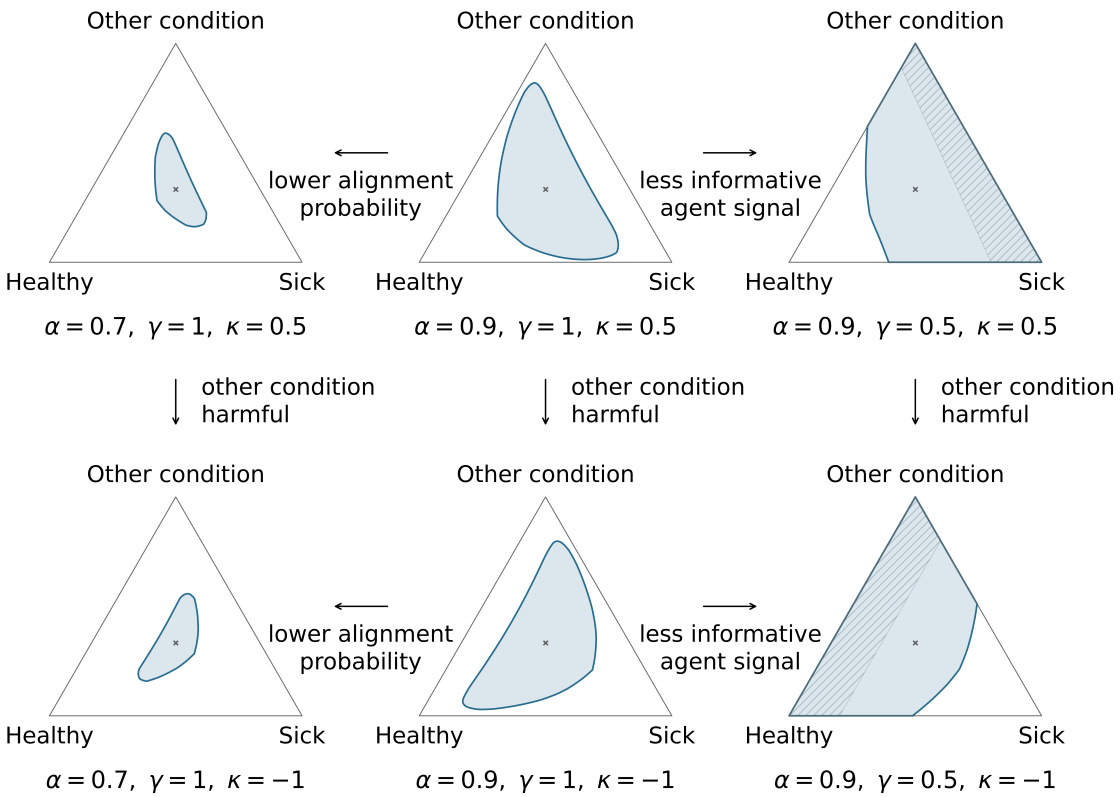


Figure 4: Comparative statics of the trust region in α, κ, and γ.

To gain economic intuition for the shape of the optimal trust region in the application, we investigate the impact of changes in key parameters, α, γ, and κ. In Figure 4, the top middle panel corresponds to the baseline parameterization. Because testing is optimal in both "Sick" and "Other" when $\kappa > 0$, the optimal trust region is skewed towards these states, with an upper boundary that is approximately parallel to the relevant boundary of the simplex—this is because it is less important for the doctor to tell apart these two states. By contrast, in the bottom middle panel, "Other" is associated with a negative payoff ($\kappa = -1$) from testing. The geometry of the trust region reflects that: The region is now skewed more towards the left upper boundary of the simplex, away from the belief that the patient is sick. This is because the doctor is now particularly concerned about misaligned AI inducing testing, by reporting that "Sick" is very likely, when it actually knows that "Other" is likely true. The effect of reducing the alignment probability, presented in the left panels in the figure, is intuitive: The trust region shrinks. Finally, the right panels in Figure 4 depict the effects of reducing the informativeness of the doctor's signal ($\gamma = 0.5$). When the doctor is less informed than AI,

she trusts it more because the relatively high alignment probability makes AI's advice highly valuable.[25] The shaded regions within the trust region correspond to interim beliefs at which the doctor ignores her own signal and always takes action "test" in the upper right panel (where $\kappa = 0.5$) and always takes action "no test" in the bottom right panel (where $\kappa = -1$). Thus, in these cases, the doctor finds it optimal to give AI authority to induce one of the actions, while retaining control over the other.

6 Concluding Remarks

Summary. We studied robust decision-making when an agent relies on an informed adviser who may be misaligned. We characterized the decision rule that maximizes the agent's expected payoff guarantee over all possible forms of misalignment. We showed that every optimal policy is equivalent to a trust region policy in belief space: the agent limits exposure to manipulation while preserving value from moderately informative advice. We proved that the optimal solution can be implemented as an equilibrium of a zero-sum game between the agent and the misaligned adviser and derived minimal alignment probabilities required for advice to be robustly valuable.

Implications for AI use. Our results support a cautiously optimistic view about deploying AI in high-stakes settings. Even if misalignment is serious and plausibly frequent, there are provably effective ways to limit the resulting harm while deriving value—provided that the human decision-maker retains final authority over actions. At the same time, our analysis makes clear that safe deployment requires concrete, pre-specified decision protocols rather than informal, case-by-case trust judgments.

The trust-region characterization translates into a simple design rule for AI-assisted choice under misalignment risk. The decision-maker should specify in advance a rule that maps model outputs into actions, separating a set of outputs that will be used directly from those that will be treated more conservatively. In some contexts, this can be implemented as

[25] While not shown in the figure, this effect is sometimes reversed at low levels of α. This is because a less informed agent is also more vulnerable to the misaligned adviser.

a delegation-style guardrail. The AI can effectively control decisions within an approved operating range, but recommendations that push beyond that range are turned into the nearest admissible recommendation or escalated into a higher-friction path (additional tests or explicit human sign-off). Moreover, when the adviser is one component of a larger AI system, the same principle suggests enforcing the trust-region mapping through a hard-coded interface layer.

Directions for future research. Our robust approach imposes no constraints on misaligned behavior. While we view this feature of our framework as well suited to the AI application, future research could identify meaningful constraints on AI behavior that can be confidently assumed. For example, if training could guarantee a minimum level of expected performance across all relevant contingencies, this guarantee could be incorporated as a constraint on the misaligned adviser's behavior. The ambiguity about misaligned behavior also contrasts with our assumption that the decision-maker has Bayesian beliefs about the AI system's capabilities (its signal structure). Relaxing this knowledge assumption would be an interesting direction for future research, since the same features of AI systems that motivate our robust approach may also make their capabilities difficult to estimate reliably. Finally, our robust approach could provide a useful lens for studying other aspects of AI deployment. For instance, there is considerable debate—among practitioners and theorists alike—about how to conceptualize and predict *interactions* among AI agents. A robust approach might help circumvent some of these difficulties by reducing the need for precise assumptions about agents' behavior, thereby contributing to the design of safer and more effective multi-agent AI systems.

References

Agarwal, N., A. Moehring, P. Rajpurkar, and T. Salz (2025): "Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology," Working paper.

Agarwal, N., A. Moehring, and A. Wolitzky (2026): "Designing Human-AI Collaboration: A Sufficient-Statistic Approach," Working paper.

Aliprantis, C. D. and K. C. Border (2006): *Infinite Dimensional Analysis: A Hitchhiker's Guide*, Berlin, Heidelberg: Springer, 3 ed.

Alonso, R., T. Gan, and J. Hu (2026): "Robust Delegation," Working paper.

Alonso, R. and G. Padró i Miquel (2025): "Competitive Capture of Public Opinion," *Econometrica*, 93, 1265–1297.

Amodei, D., C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané (2016): "Concrete Problems in AI Safety," Working paper.

Benjamin, D. J. (2019): "Errors in Probabilistic Reasoning and Judgment Biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2, 69–186.

Betley, J., N. Warncke, A. Sztyber-Betley, D. Tan, X. Bao, M. Soto, M. Srivastava, N. Labenz, and O. Evans (2026): "Training large language models on narrow tasks can lead to broad misalignment," *Nature*, 649, 584–589.

Blackwell, D. (1951): "Comparison of Experiments," in *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, ed. by J. Neyman, Berkeley and Los Angeles: University of California Press, 93–102.

Bregman, L. M. (1967): "The Relaxation Method of Finding the Common Point of Convex Sets and Its Application to the Solution of Problems in Convex Programming," *USSR Computational Mathematics and Mathematical Physics*, 7, 200–217.

Chen, E. O., A. Ghersengorin, and S. Petersen (2026): "AI Control and Imperfect Recall," Working paper.

Chen, Y. (2011): "Perturbed Communication Games with Honest Senders and Naive Receivers," *Journal of Economic Theory*, 146, 401–424.

Crawford, V. P. and J. Sobel (1982): "Strategic Information Transmission," *Econometrica*, 50, 1431–51.

D'Amour, A., K. Heller, D. Moldovan, B. Adlam, B. Alipanahi, A. Beutel, C. Chen, J. Deaton, J. Eisenstein, M. D. Hoffman, F. Hormozdiari, N. Houlsby, S. Hou, G. Jerfel, A. Karthikesalingam, M. Lucic, Y. Ma, C. McLean, D. Mincu, A. Mitani, A. Montanari, Z. Nado, V. Natarajan,

C. Nielson, T. F. Osborne, R. Raman, K. Ramasamy, R. Sayres, J. Schrouff, M. Seneviratne, S. Sequeira, H. Suresh, V. Veitch, M. Vladymyrov, X. Wang, K. Webster, S. Yadlowsky, T. Yun, X. Zhai, and D. Sculley (2022): "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *Journal of Machine Learning Research*, 23, 1–61.

Dilmé, F. (2022): "Strategic Communication with a Small Conflict of Interest," *Games and Economic Behavior*, 134, 1–19.

Doval, L. and A. Smolin (2024): "Persuasion and Welfare," *Journal of Political Economy*, 132, 2451–2487.

Dratsch, T., X. Chen, M. Rezazade Mehrizi, R. Kloeckner, A. Mähringer-Kunz, M. Püsken, B. Baessler, S. Sauer, D. Maintz, and D. Pinto dos Santos (2023): "Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance," *Radiology*, 307, e222176.

Dreyfuss, B. and R. Hoong (2025): "Calibrated Coarsening: Designing Information for AI-Assisted Decisions," Working paper.

Dworczak, P. and A. Pavan (2022): "Preparing for the Worst but Hoping for the Best: Robust (Bayesian) Persuasion," *Econometrica*, 90, 2017–2051.

Dziuda, W. (2011): "Strategic Argumentation," *Journal of Economic Theory*, 146, 1362–1397.

Frankel, A. (2014): "Aligned Delegation," *American Economic Review*, 104, 66–83.

Fudenberg, D. and A. Liang (2026): "Friend or Foe: Delegating to an AI whose Alignment is Unknown," Working paper.

Gale, D. and H. Nikaido (1965): "The Jacobian Matrix and Global Univalence of Mappings," *Mathematische Annalen*, 159, 81–93.

Gershkov, A., B. Moldovanu, and X. Shi (2025): "Order Independence in Sequential, Issue-by-Issue Voting," *Mathematics of Operations Research*, 50, 1635–1653.

Glazer, J., H. Herrera, and M. Perry (2021): "Fake Reviews," *The Economic Journal*, 131, 1772–1787.

Hurwicz, L. (1951): "Optimality Criteria for Decision Making Under Ignorance," *Cowles Commission Discussion Paper*.

Jacod, J. and J. Mémin (1981): "Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité," *Séminaire de probabilités*, 15, 529–546.

——— (1983): "Rectification à "Sur un type de convergence intermédiaire entre la convergence en loi et la convergence en probabilité"," *Séminaire de probabilités*, 17, 509–511.

Kalai, A. T. and S. S. Vempala (2024): "Calibrated Language Models Must Hallucinate," in *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, 160–171.

Kallenberg, O. (2021): *Foundations of Modern Probability*, vol. 99 of *Probability Theory and Stochastic Modelling*, Cham: Springer, 3 ed.

Kamenica, E. and M. Gentzkow (2011): "Bayesian Persuasion," *American Economic Review*, 101, 2590–2615.

Lahr, P. and J. Winkelmann (2019): "Fake Experts," Working paper.

Langosco, L. L. D., J. Koch, L. D. Sharkey, J. Pfau, and D. Krueger (2022): "Goal Misgeneralization in Deep Reinforcement Learning," in *Proceedings of the 39th International Conference on Machine Learning*, PMLR, vol. 162 of *Proceedings of Machine Learning Research*, 12004–12019.

Levy, M. and B. Szentes (2025): "Information Design for AI Proxies under Imperfect Recall," Working paper.

Lipnowski, E. and D. Ravid (2020): "Cheap Talk with Transparent Motives," *Econometrica*, 88, 1631–1660.

Lipnowski, E., D. Ravid, and D. Shishkin (2022): "Persuasion via Weak Institutions," *Journal of Political Economy*, 130, 2705–2730.

Maslej, N., L. Fattorini, R. Perrault, Y. Gil, V. Parli, N. Kariuki, E. Capstick, A. Reuel, E. Brynjolfsson, J. Etchemendy, K. Ligett, T. Lyons, J. Manyika, J. C. Niebles, Y. Shoham, R. Wald, T. Walsh, A. Hamrah, L. Santarlasci, J. B. Lotufo, A. Rome, A. Shi, and S. Oak (2025): "The AI Index 2025 Annual Report," Tech. rep., AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA.

Milgrom, P. and I. Segal (2002): "Envelope Theorems for Arbitrary Choice Sets," *Econometrica*, 70, 583–601.

Min, D. (2021): "Bayesian Persuasion under Partial Commitment," *Economic Theory*, 72, 743–764.

Russell, S. (2019): *Human Compatible: AI and the Problem of Control*, Penguin UK.

Sharma, M., M. Tong, T. Korbak, D. Duvenaud, A. Askell, S. R. Bowman, E. Durmus, Z. Hatfield-Dodds, S. R. Johnston, S. M. Kravec, T. Maxwell, S. McCandlish, K. Ndousse, O. Rausch, N. Schiefer, D. Yan, M. Zhang, and E. Perez (2024): "Towards Understanding Sycophancy in Language Models," in *International Conference on Learning Representations (ICLR)*.

Sion, M. (1958): "On General Minimax Theorems," *Pacific Journal of Mathematics*, 8, 171–176.

Sobel, J. (2020): "Lying and Deception in Games," *Journal of Political Economy*, 128, 907–947.

Wenderott, K., J. Krups, F. Zaruchas, and M. Weigl (2024): "Effects of artificial intelligence implementation on efficiency in medical imaging—a systematic literature review and meta-analysis," *npj Digital Medicine*, 7, 265.

Whitmeyer, M. (2026): "Blackwell-Monotone Updating Rules," *Journal of Political Economy*, forthcoming.

Wojtaszczyk, P. (1991): *Banach Spaces for Analysts*, vol. 25 of *Cambridge Studies in Advanced Mathematics*, Cambridge, UK: Cambridge University Press.

A Proofs

A.1 Proofs of Theorem 1 and Theorem 2

Both the proof of Theorem 1 and the proof of Theorem 2 represent the agent's strategies as mappings from messages to state-contingent payoff profiles (cf. Doval and Smolin (2024)). We first collect a few key lemmas that are needed for both proofs, and then complete the arguments in Appendices A.2 and A.3, respectively.

For a private strategy $\hat{\sigma}$, define

$$\mathrm{w}(\hat{\sigma})(\omega) \triangleq \mathbb{E}_{\hat{\sigma}}[u(a,\omega,\theta) \mid \omega] = \int_\Theta \int_A u(a,\omega,\theta)\hat{\sigma}(da \mid \theta) f(d\theta \mid \omega), \qquad \omega \in \Omega,$$

and let $W = \{\mathrm{w}(\hat{\sigma}) : \hat{\sigma} \text{ is a private strategy}\}$ be the set of feasible payoff profiles. If the agent holds belief $\mu \in \Delta(\Omega)$ over the state and chooses a private strategy $\hat{\sigma}$, her expected payoff is $\mu \cdot \mathrm{w}(\hat{\sigma})$. Conditional independence of θ and s given ω ensures that the agent's expected payoff conditional on the adviser's belief s is likewise $s \cdot \mathrm{w}(\hat{\sigma})$.

We first establish basic properties of the set W and a formal equivalence between strategies in the original problem and mappings from messages to the set W.

Lemma 2. *The set W is compact and convex. Every strategy $\sigma \sim (\hat{\sigma}(m))_{m\in\Delta(\Omega)}$ induces a measurable payoff mapping $g_\sigma : \Delta(\Omega) \to W$, given by $g_\sigma(m) = \mathrm{w}(\hat{\sigma}(m))$. Conversely, every measurable mapping $g : \Delta(\Omega) \to W$ can be implemented by a strategy $\sigma \sim (\hat{\sigma}(m))_{m\in\Delta(\Omega)}$ such that $\mathrm{w}(\hat{\sigma}(m)) = g(m)$ for every $m \in \Delta(\Omega)$.*

Proof. Let $\bar{f} = \sum_\omega \mu_0(\omega) f(\cdot \mid \omega)$ and $r_\omega = df(\cdot \mid \omega)/d\bar{f}$; the Radon–Nikodym derivatives exist because μ_0 has full support. Every private strategy $\hat{\sigma}$ induces a joint distribution $\rho(d\theta, da) = \bar{f}(d\theta)\hat{\sigma}(da \mid \theta)$ in $\mathcal{R} \triangleq \{\rho \in \Delta(\Theta \times A) : \rho_\Theta = \bar{f}\}$. Conversely, the disintegration theorem gives, for every $\rho \in \mathcal{R}$, a probability kernel $\hat{\sigma}(da \mid \theta)$ with $\rho(d\theta, da) = \bar{f}(d\theta)\hat{\sigma}(da \mid \theta)$. The payoff profile induced by ρ is given by

$$\Phi(\rho)(\omega) \triangleq \int_{\Theta\times A} r_\omega(\theta) u(a,\omega,\theta)\, \rho(d\theta, da).$$

Thus, $W = \Phi(\mathcal{R})$. The set $\mathcal{R}$ is compact and convex in the weak topology, and Φ is affine and continuous.[26] Thus W is compact and convex.

For any strategy σ, the function $m \mapsto \mathrm{w}(\hat{\sigma}(m))(\omega)$ is measurable for each ω, by integration of the measurable kernel σ against the fixed conditional type distribution $f(\cdot \mid \omega)$. Conversely, for any measurable g, the measurable maximum theorem (Theorem 18.19 in Aliprantis and

[26] To establish continuity, we can first apply the fixed-marginal convergence result of Jacod and Mémin (1981, 1983) to the integrand $r_\omega(\theta)u(a,\omega,\theta)$ truncated to $[-K, K]$. Since $r_\omega(\theta)\sup_a |u(a,\omega,\theta)|$ is integrable under $\bar{f}$ (by the uniform integrability assumption on u), and every $\rho \in \mathcal{R}$ has the same Θ-marginal $\bar{f}$, the truncation error is bounded uniformly over $\mathcal{R}$ by $\int (r_\omega(\theta)\sup_a |u(a,\omega,\theta)| - K)_+ \, d\bar{f}$, which vanishes as $K \to \infty$. Thus, the original integral is a uniform limit of continuous functions.

Border (2006)) applied to $-\|\Phi(\rho) - g(m)\|$ yields a measurable selection $m \mapsto \rho_m \in \mathcal{R}$ satisfying $\Phi(\rho_m) = g(m)$. The parameterized disintegration theorem (Theorem 9.27 in Kallenberg (2021)), applied to the probability kernel $m \mapsto \rho_m$ with common Θ-marginal $\bar{f}$, yields a probability kernel $\sigma(da \mid m, \theta)$ measurable in (m, θ) and satisfying $\rho_m(d\theta, da) = \bar{f}(d\theta)\sigma(da \mid m, \theta)$ for every m. Thus σ implements g. □

Remark 2. There is a close connection between Bayes-optimal private strategies and the Pareto frontier of the set W. A profile $\mathrm{w} \in W$ is induced by a Bayes-optimal private strategy at some belief if and only if it is *weakly Pareto optimal* in W: there is no $v \in W$ with $v(\omega) > \mathrm{w}(\omega)$ for every $\omega \in \Omega$. In one direction, a Bayes-optimal profile cannot be strictly improved in every state. For the converse, if w is weakly Pareto optimal, the convex sets W and $\mathrm{w} + \mathbb{R}^N_{++}$ are disjoint. Since W is compact and convex by Lemma 2, the separating-hyperplane theorem yields a nonzero vector $p \in \mathbb{R}^N_+$ such that $p \cdot \mathrm{w} \geq p \cdot v$ for all $v \in W$. Normalizing p to sum to one yields a belief at which w is Bayes-optimal.

For the existence proof, it is helpful to further simplify the problem by observing that the agent's overall strategy determines a set of available payoff profiles (which we will refer to as a *menu*) and a matching of messages to these profiles. For a fixed menu, the matching can be optimized as if the adviser were always aligned: each message is assigned a payoff profile that maximizes the agent's expected payoff at the corresponding belief. In contrast, when the adviser is misaligned, the worst-case payoff depends only on the menu, and is given by the minimum expected payoff over its elements. We formalize these intuitions next.

Let $\mathcal{K}_c(W)$ be the set of nonempty compact convex subsets of W. For $C \in \mathcal{K}_c(W)$, let $h_C(m) = \max_{\mathrm{w} \in C} m \cdot \mathrm{w}$ and $\ell_C(m) = \min_{\mathrm{w} \in C} m \cdot \mathrm{w}$, and define

$$G(C) \triangleq \int_M \big[\alpha h_C(s) + (1-\alpha)\ell_C(s)\big]\tau(ds).$$

For a measurable mapping $g : \Delta(\Omega) \to W$ and a reporting strategy β, we will abuse notation slightly by writing

$$V(\beta, g) = \alpha \int_M s \cdot g(s)\, \tau(ds) + (1-\alpha) \int_M \int_{\Delta(\Omega)} s \cdot g(m)\, \beta(dm \mid s)\tau(ds),$$

noting that this definition agrees with the definition of $V(\beta, \sigma)$ whenever $g = g_\sigma$ (see Lemma 2). Similarly, we will write $V(g) = \inf_{\beta \in \mathcal{B}} V(\beta, g)$. Finally, let $K_g = \overline{\text{co}}\{g(m) : m \in \Delta(\Omega)\}$, where $\overline{\text{co}}$ denotes the closed convex hull operator.

Lemma 3. *For every measurable mapping $g : \Delta(\Omega) \to W$,*

$$V(g) = \int_M \big[\alpha s \cdot g(s) + (1-\alpha)\ell_{K_g}(s)\big]\tau(ds).$$

Moreover, $V^* = \sup_{g:\Delta(\Omega)\to W} V(g) = \max_{C \in \mathcal{K}_c(W)} G(C)$, and if C^* maximizes G, then any measurable selection $g^*(m) \in \arg\max_{\mathrm{w} \in C^*} m \cdot \mathrm{w}$ corresponds to an optimal strategy.

Proof. Since a linear functional has the same infimum over a set and its closed convex hull, $\inf_{m \in \Delta(\Omega)} s \cdot g(m) = \ell_{K_g}(s)$. Thus, $V(g)$ is bounded below by $\int_M \big[\alpha s \cdot g(s) + (1-\alpha)\ell_{K_g}(s)\big]\tau(ds)$. The reverse inequality also holds, because the misaligned adviser can approximate this bound using measurable reporting strategies.[27]

The proven expression for $V(g)$ immediately implies that $V(g) \leq G(K_g)$. Conversely, for any menu C, the measurable maximum theorem gives a measurable selection $g_C(m) \in \arg\max_{\mathrm{w} \in C} m \cdot \mathrm{w}$, implemented by Lemma 2. When the adviser's true belief is s, this mapping yields $h_C(s)$ under truthful reporting and at least $\ell_C(s)$ under any reporting strategy of the misaligned adviser. Hence, $V(g_C) \geq G(C)$. Therefore $V^* = \sup_{C \in \mathcal{K}_c(W)} G(C)$.

The space $\mathcal{K}_c(W)$ is compact under Hausdorff distance: nonempty compact subsets of W form a compact space (Theorem 3.85(3) in Aliprantis and Border (2006)), and convexity is preserved under Hausdorff limits. Both functions $h_C(m)$ and $\ell_C(m)$ change by at most $d_H(C, D)$ when C is replaced by D, since $\|m\| \leq 1$ on the space of beliefs. Hence, G is continuous and has a maximizer C^*. By the equality we established above, any maximizing selection g^* from C^* guarantees V^* and is therefore optimal. □

The final preliminary lemma proves that there exists a common minimax reporting strategy for all optimal strategies of the agent. This lemma is a key ingredient of the proof of

[27] Formally, we can choose messages $(m_j)_{j \geq 1}$ such that $\{g(m_j) : j \geq 1\}$ is dense in the image of g. For each k, let $j_k(s)$ be the smallest index minimizing $s \cdot g(m_j)$ over $1 \leq j \leq k$. Since this minimum is taken over finitely many continuous functions of s, the mapping $s \mapsto j_k(s)$ is measurable, and hence $\beta_k(\cdot \mid s) = \delta_{m_{j_k(s)}}$ defines a reporting strategy. As $k \to \infty$, the expected payoff at each belief s converges to $\ell_{K_g}(s)$. Bounded convergence then implies that the misaligned adviser's expected payoff under β_k converges to $\int_{\Delta(\Omega)} \ell_{K_g}(s)\tau(ds)$.

the minimax result in Theorem 2 but it is also used in the proof of Theorem 1.

Lemma 4. *There exists a reporting strategy $\beta^* \in \mathcal{B}$ such that $V(\beta^*, \sigma) \leq V^*$ for every $\sigma \in \Sigma$. For any optimal strategy σ^*, (β^*, σ^*) is a saddle point:*

$$V(\beta^*, \sigma) \leq V(\beta^*, \sigma^*) \leq V(\beta, \sigma^*) \qquad \text{for every } \sigma \in \Sigma,\ \beta \in \mathcal{B}. \tag{11}$$

Proof. We first represent the adviser's reporting strategy by a joint distribution of the true belief and message in the compact convex set

$$\Gamma = \{\chi \in \Delta(M \times M) : \chi_1 = \tau\},$$

where χ_1 and χ_2 denote the marginal distributions of true beliefs and messages under χ, respectively. Every $\chi \in \Gamma$ corresponds to a reporting strategy $\beta \in \mathcal{B}$ supported on M, since the disintegration theorem guarantees a probability kernel satisfying $\chi(ds, dm) = \tau(ds)\beta(dm \mid s)$. We will show that restricting attention to such reporting strategies is sufficient to construct a common minimax strategy. Abusing notation again, for measurable $g : \Delta(\Omega) \to W$, we write

$$V(\chi, g) = \alpha \int_M s \cdot g(s)\, \tau(ds) + (1 - \alpha) \int_{M \times M} s \cdot g(m)\, \chi(ds, dm),$$

which agrees with $V(\beta, g)$ when χ is the joint distribution induced by β.

On the space $\Gamma \times \mathcal{C}(M, W)$, where $\mathcal{C}(M, W)$ is the convex set of continuous mappings from M to W, the function $V(\chi, g)$ is affine in each argument, continuous in χ under the weak topology for fixed g, and continuous in g under the uniform topology for fixed χ. The minimax theorem of Sion (1958) therefore yields

$$v^* := \sup_{g \in \mathcal{C}(M,W)} \min_{\chi \in \Gamma} V(\chi, g) = \min_{\chi \in \Gamma} \sup_{g \in \mathcal{C}(M,W)} V(\chi, g).$$

The minimum on the right-hand side is attained at some $\chi^* \in \Gamma$, since the supremum of continuous functions of χ is lower semicontinuous and Γ is compact.

To extend the minimax inequality from continuous to all measurable mappings g, we now fix χ^* and use approximation in the $L^1(q^*)$ topology, where $q^* = \alpha\tau + (1 - \alpha)\chi_2^*$ is the

unconditional distribution of messages when the misaligned adviser uses χ^*. Every measurable $g : M \to W$ is an $L^1(q^*)$ limit of continuous maps into W.[28] Since $\|s\| \leq 1$ for every belief s, we have

$$|V(\chi^*, g) - V(\chi^*, h)| \leq \int_M \|g(m) - h(m)\| \, q^*(dm).$$

Thus, the minimax inequality extends to every measurable payoff mapping g:

$$V(\chi^*, g) \leq v^* \qquad \text{for every measurable } g : M \to W. \tag{12}$$

Let $\beta^*(dm \mid s)$ be a reporting strategy satisfying $\chi^*(ds, dm) = \tau(ds)\beta^*(dm \mid s)$. By Lemma 2, inequality (12) applies to the payoff mapping induced by every agent strategy, implying $V^* \leq v^*$. Conversely, for each continuous $g \in \mathcal{C}(M, W)$, let $C = \overline{\text{co}}\{g(m) : m \in M\}$. As in Lemma 3,

$$\min_{\chi \in \Gamma} V(\chi, g) = \int_M \left[\alpha s \cdot g(s) + (1 - \alpha)\ell_C(s)\right] \tau(ds) \leq G(C) \leq V^*.$$

Thus, $v^* = V^*$, and $V(\beta^*, \sigma) \leq V^*$ for every agent strategy σ. Finally, optimality of σ^* gives $V(\beta, \sigma^*) \geq V^*$ for every β. Combining these two bounds establishes (11). □

A.2 Completion of proof of Theorem 1

Existence of an optimal strategy follows from Lemma 2 and Lemma 3.

Throughout the proof, we focus on the nontrivial case $\alpha > 0$.[29] Fix an optimal strategy σ^*. The proof that σ^* is equivalent to a TRS with a connected trust region proceeds in three steps. First, we enlarge the menu of payoff profiles so that its maximizing profiles are Bayes-optimal at some belief, and construct an equivalent strategy that optimally maps messages to these

[28] Approximate each coordinate by continuous functions (Theorems 12.7 and 13.9 in Aliprantis and Border (2006)), then use the fact that Euclidean projection onto the compact convex set W is 1-Lipschitz.

[29] If $\alpha = 0$, the agent can attain the optimal value $U(\mu_0)$ by ignoring the adviser, while the misaligned adviser can guarantee the corresponding upper bound by sending any fixed message. In particular, for an optimal strategy and any message m, sending m with probability one gives the agent payoff $\mu_0 \cdot g(m)$, so optimality requires $\mu_0 \cdot g(m) = U(\mu_0)$. Every response is therefore Bayes-optimal at μ_0. Fixing any message m_0, the constant strategy that uses $\hat{\sigma}^*(m_0)$ after every message is a TRS with $T = \{\mu_0\}$ and $P(m) = \mu_0$, equivalent to the original strategy under the adviser who always reports m_0.

profiles. Second, we construct the corresponding trust region in the belief space. Finally, we show that responses outside the trust region can be implemented by projecting messages onto its visible boundary.

Step 1: Enlarge the menu of payoff profiles. Let $g(m) = \mathrm{w}(\hat{\sigma}^*(m))$ be the mapping from messages to state-contingent payoff profiles corresponding to the optimal strategy σ^* (as in Lemma 2), $K_g = \overline{\mathrm{co}}\{g(m) : m \in \Delta(\Omega)\}$, and

$$C = (K_g + \mathbb{R}^N_+) \cap W.$$

The set C is compact, convex, and upward closed within W. Adding payoff profiles that weakly improve on an existing profile in every state does not lower the worst payoff at any belief: $\ell_C = \ell_{K_g}$, while $h_C \geq h_{K_g}$ because $C \supset K_g$. Hence, Lemma 3 gives

$$V^* = V(g) \leq G(K_g) \leq G(C) \leq V^*.$$

The worst-case payoff against the misaligned adviser depends on the available payoff profiles, rather than their assignment to messages. Holding C fixed, the agent can therefore maximize the expected payoff at each message as if the adviser were aligned (i.e., treating the message as the correct interim belief). Formally, the equality above gives

$$0 = G(C) - V(g) = \alpha \int_M \left[h_C(s) - s \cdot g(s)\right] \tau(ds).$$

Since $\alpha > 0$ and the integrand is nonnegative, the set $E = \{m \in \Delta(\Omega) : m \cdot g(m) = h_C(m)\}$ satisfies $\tau(E) = 1$. Choose a measurable mapping $\overline{g} : \Delta(\Omega) \to C$ that equals g on E and satisfies $\overline{g}(m) \in \arg\max_{\mathrm{w} \in C} m \cdot \mathrm{w}$ for every $m \in \Delta(\Omega)$. Such a mapping exists by the measurable maximum theorem. Retain the original private strategies $\hat{\sigma}^*(m)$ on E and implement $\overline{g}$ elsewhere using Lemma 2. The resulting strategy $\overline{\sigma}$ is optimal by Lemma 3.

Although g and $\overline{g}$ coincide τ-almost surely, they may induce different outcomes when the adviser is misaligned, since he can assign positive probability to messages in a τ-null set. To establish equivalence of the original and modified agent strategies in terms of the induced outcome distributions, we construct a common minimax reporting strategy whose

induced distribution of messages is absolutely continuous with respect to τ. Not every optimal reporting strategy has this property. Intuitively, however, messages in a τ-null set reveal that the adviser is misaligned, allowing the agent to respond to them independently of truthful communication. This allows us to modify a common minimax reporting strategy without compromising its minimax property.

Let β^* be the common minimax strategy from Lemma 4, and denote by

$$\nu^*(A) = \int_M \beta^*(A \mid s)\tau(ds)$$

the marginal distribution of messages under the joint distribution $\tau(ds)\beta^*(dm \mid s)$, for every measurable set $A \subseteq M$. By the Lebesgue decomposition theorem, there is a measurable set $Z \subseteq M$ such that $\tau(Z) = 0$ and the restriction of ν^* to $M \setminus Z$ is absolutely continuous with respect to τ. Define a reporting strategy $\beta^{**} : M \to \Delta(M)$ by

$$\beta^{**}(A \mid s) = \beta^*(A \setminus Z \mid s) + \beta^*(Z \mid s)\tau(A)$$

for every Borel set $A \subseteq M$. Thus, whenever the original strategy would send a message in Z, that message is replaced by a draw from τ, independently of the true belief s. The marginal distribution of messages under $\tau(ds)\beta^{**}(dm \mid s)$ is $\nu^{**}(A) = \nu^*(A \setminus Z) + \nu^*(Z)\tau(A)$, so ν^{**} is absolutely continuous with respect to τ.

Moreover, β^{**} remains a common minimax reporting strategy. To see this, fix any measurable $h : \Delta(\Omega) \to W$, let $\bar{w} = \int_M h(t)\tau(dt) \in W$, and define $\overline{h} : \Delta(\Omega) \to W$ to equal $\bar{w}$ on Z and h elsewhere. For each $s \in M$,

$$\int_M s \cdot h(m)\, \beta^{**}(dm \mid s) = \int_{M \setminus Z} s \cdot h(m)\, \beta^*(dm \mid s) + \beta^*(Z \mid s)\, s \cdot \bar{w} = \int_M s \cdot \overline{h}(m)\, \beta^*(dm \mid s).$$

The aligned-adviser contributions also agree because $\tau(Z) = 0$. Consequently,

$$V(\beta^{**}, h) = V(\beta^*, \overline{h}) \leq V^*.$$

The last inequality follows from Lemma 4, since $\overline{h}$ is implementable by Lemma 2. Thus, β^{**}

remains a common minimax reporting strategy and is adversarial to every optimal agent strategy, including σ^* and $\overline{\sigma}$.

Accounting for both adviser types, the unconditional distribution of messages under β^{**} is

$$q^* = \alpha\tau + (1-\alpha)\nu^{**} \ll \tau.$$

Since σ^* and $\overline{\sigma}$ use the same private strategies on E and $q^*(E) = 1$, they induce the same joint distribution of states, types, messages, and actions against β^{**}. Thus, σ^* and $\overline{\sigma}$ are equivalent.

Step 2: Construct a compact, connected trust region. Fixing the equivalent optimal strategy $\overline{\sigma}$ and the (compact and convex) menu C constructed in step 1, define $F(m) = \arg\max_{\mathrm{w}\in C} m\cdot \mathrm{w}$, and let $D = \bigcup_{m\in\Delta(\Omega)} F(m) \subseteq C$ be its image. Intuitively, D is the subset of payoff profiles of C that are an optimal response for the agent for some message m (treating that message as the correct interim belief). The correspondence F has a compact graph and nonempty convex values over the connected domain $\Delta(\Omega)$. Consequently, its graph is connected, and so is its image D.[30] Since C is upward closed, every $\mathrm{w} \in D$ is weakly Pareto optimal in W. Indeed, if $v \in W$ strictly improved on w in every state, then $v \in C$; but at any belief m for which $\mathrm{w} \in \arg\max_{w'\in C} m \cdot w'$, we would have $m \cdot v > m \cdot \mathrm{w}$, a contradiction.

Since every payoff profile in D is weakly Pareto optimal in W, it is Bayes-optimal at some belief, by Remark 2. For each $\mathrm{w} \in D$, let

$$B(\mathrm{w}) = \{\mu \in \Delta(\Omega) : \mu \cdot \mathrm{w} = U(\mu)\}$$

denote the set of beliefs at which w is Bayes-optimal. This set is nonempty, compact, and convex. Moreover, the correspondence B has a compact graph over the connected set D. By the same argument as above, its graph and image are connected. We can therefore define the

[30] A compact graph with connected values over a connected domain is connected: otherwise, it could be partitioned into two disjoint nonempty closed sets whose projections would likewise partition the domain into two disjoint nonempty closed sets.

nonempty, compact, and connected trust region

$$T = \bigcup_{\mathrm{w} \in D} B(\mathrm{w}) = \{\mu \in \Delta(\Omega) : h_C(\mu) = U(\mu)\}.$$

By construction, $h_C(\mu) = U(\mu)$ for every $\mu \in T$, so any payoff profile maximizing $\mu \cdot \mathrm{w}$ over C is also Bayes-optimal at μ. Thus, $\overline{\sigma}$ assigns a Bayes-optimal private strategy to every message $m \in T$, satisfying the first requirement of a TRS.

Step 3: Map untrusted messages to the visible boundary. For every $m \notin T$, choose a measurable mapping $\overline{\mu}$ satisfying $\overline{\mu}(m) \in B(\overline{g}(m))$. Such a mapping exists by the measurable maximum theorem applied to the function $(m, \mu) \mapsto \mu \cdot \overline{g}(m) - U(\mu)$ on $\Delta(\Omega) \times \Delta(\Omega)$. The payoff profile $\overline{g}(m)$ maximizes expected payoff over C at m by construction, and at $\overline{\mu}(m)$ because it is Bayes-optimal there. By linearity of expected payoffs in beliefs, the same profile maximizes expected payoff over C at every belief on the segment joining m and $\overline{\mu}(m)$.

For every $m \notin T$, let $P(m)$ be the first point of T encountered when moving along the segment from m toward $\overline{\mu}(m)$, and set $P(m) = m$ for $m \in T$. This projection is well-defined because T is compact and $\overline{\mu}(m) \in T$.[31] By construction, $P(m)$ is visible from m. Moreover, the preceding argument shows that $\overline{g}(m)$ maximizes expected payoff over C at $P(m)$. Since $P(m) \in T$, the private strategy prescribed by $\overline{\sigma}$ after message m is Bayes-optimal at $P(m)$. Thus, $\overline{\sigma}$ is a TRS.

Proof of Corollary 1. The singleton trust region for $\alpha = 0$ is non-hollow. For $\alpha > 0$, take $\mu \in \Delta(\Omega) \setminus T$ and a Bayes-optimal profile w at μ. Then $\mathrm{w} \notin C$. Since $\mathrm{w} \in W$ and C is upward closed within W, we also have $\mathrm{w} \notin C + \mathbb{R}^N_+$. Separating w from the closed convex set $C + \mathbb{R}^N_+$ yields a nonzero nonnegative vector, which can be normalized to a belief $s \in \Delta(\Omega)$ satisfying $s \cdot \mathrm{w} < \ell_C(s)$.

Consider any $q = \mu + \lambda(\mu - s) \in \Delta(\Omega)$, $\lambda > 0$, and any Bayes-optimal profile v at q.

[31] To verify measurability on $\Delta(\Omega) \setminus T$, let $\lambda_*(m)$ be the smallest $\lambda \in [0, 1]$ such that $(1-\lambda)m + \lambda\overline{\mu}(m) \in T$. For any $r \in [0, 1]$, the condition $\lambda_*(m) \leq r$ is equivalent to $\min_{\lambda \in [0,r]} \mathrm{dist}((1-\lambda)m + \lambda\overline{\mu}(m), T) = 0$. This minimum is measurable in m, since it equals the infimum over a countable dense subset of $[0, r]$ of measurable functions. Hence λ_* and P are measurable.

Optimality gives

$$\mu \cdot (v - \mathrm{w}) \leq 0 \leq q \cdot (v - \mathrm{w}) = (1 + \lambda)\mu \cdot (v - \mathrm{w}) - \lambda s \cdot (v - \mathrm{w}).$$

Thus $s \cdot v \leq s \cdot \mathrm{w} < \ell_C(s)$, so no Bayes-optimal profile at q belongs to C, implying $q \notin T$. Therefore, starting from any $\mu \notin T$, we can move along the ray pointing away from s without encountering T, establishing non-hollowness.

Proof of Corollary 2. Fix any optimal strategy, and let σ be the equivalent optimal TRS constructed in the proof of Theorem 1. Suppose first that $\alpha > 0$. By construction, $\hat{\sigma}(m)$ maximizes the agent's payoff at m over the menu C, which contains the payoff profiles of all private strategies in $\mathrm{Im}(\sigma)$. Moreover, for every $m \notin T$, $\hat{\sigma}(m)$ is Bayes-optimal at $P(m)$. Thus,

$$d_\sigma(m, P(m)) = U(m) - U(\hat{\sigma}(m), m) \leq d_\sigma(m, \mu) \qquad \text{for every } \mu \in T.$$

For $\alpha = 0$, the result follows immediately from the singleton trust region construction in the proof of Theorem 1.

A.3 Completion of the proof of Theorem 2

By Lemma 3 and Theorem 1, there exists an optimal TRS σ^* with trust region T and measurable map P. Pairing it with the adversarial reporting strategy β^* from Lemma 4, we obtain a saddle point (β^*, σ^*). It remains to establish that σ^* prescribes Bayes-optimal responses to the posterior beliefs induced by β^*.

Let

$$q^*(A) = \alpha\tau(A) + (1 - \alpha) \int_M \beta^*(A \mid s)\tau(ds), \qquad A \subseteq M,$$

be the resulting distribution of messages. For each $\omega \in \Omega$, define the measure

$$\nu_\omega(A) = \alpha \int_A s(\omega)\tau(ds) + (1 - \alpha) \int_M s(\omega)\beta^*(A \mid s)\tau(ds), \qquad A \subseteq M.$$

Since $\sum_{\omega \in \Omega} \nu_\omega = q^*$, the Radon–Nikodym derivatives $p_\omega = d\nu_\omega / dq^*$ exist and define a measurable posterior $p(m) = (p_\omega(m))_{\omega \in \Omega} \in \Delta(\Omega)$ for q^*-almost every m.

By inequality (11), σ^* is a best response to β^*. Since every measurable payoff mapping is implementable by Lemma 2, the agent's prescribed response must maximize expected payoff at the posterior belief $p(m)$ for q^*-almost every message m. Thus,

$$p(m) \cdot \mathrm{w}(\hat{\sigma}^*(m)) = U(p(m)) \qquad q^*\text{-almost surely.}$$

Let $D \subseteq M$ be the measurable set on which this equality holds. Then $q^*(D) = 1$.

At each remaining message $m \notin D$, define the posterior belief to be m if $m \in T$ and $P(m)$ otherwise. These beliefs rationalize the prescribed responses because σ^* is a TRS. Since $q^*(D) = 1$, this extension of the posterior mapping leaves the joint distribution of states and messages unchanged. Thus σ^* is robustly rationalizable.

A.4 Proof of Theorem 3

We first establish the bounds and attainment result for finite M, with $|M| = K$. We then extend the bounds to arbitrary supports.

Notation. In this appendix, we denote by I_K a unit matrix of dimension K, by 1_K a vector of ones of dimension K, by $0_{N\times K}$ a matrix of zeros of dimension $N \times K$, by e_K^i an ith standard basis vector of dimension K, by $x_{i,k}$ a kth element of vector x_i, by $\operatorname{diag} x$ a diagonal matrix with vector x on the main diagonal, by δ_{ij} the Kronecker delta (equal to 1 if $i = j$ and 0 otherwise), and by $X^\top$ a transpose of a matrix X.

Since μ_0 has full support, we can equivalently identify the adviser's information with a (row) stochastic $N \times K$ matrix Π, where Π_{ij} is the probability of the jth signal observed by the adviser in the ith state. Moreover, $\operatorname{rank}\Pi = R(\tau)$.[32]

We identify the strategy of the misaligned adviser with a stochastic $K \times K$ matrix B. Since the aligned adviser reports truthfully, the overall adviser's strategy can be written as a

[32] A matrix of adviser's posteriors can be computed by Bayes' rule as $(\mu(s))_{s\in S} = (\operatorname{diag}(\mu_0(\omega))_{\omega\in\Omega})\Pi(\operatorname{diag}(\tau(s))_{s\in S})^{-1}$. The diagonal matrices are invertible and the multiplication by an invertible matrix preserves the rank.

garbling of his information:

$$G(B) \triangleq \alpha I_K + (1-\alpha)B. \tag{13}$$

By Blackwell (1951), the MVA is a maximal α for which there exists a stochastic matrix B such that $\Pi G(B)$ is Blackwell uninformative; we show below that it is attained. This also implies that MVA depends on τ only via Π, so we will write MVA(Π).

We start with preliminary observations. First, note that $G_{kk} \geq \alpha$ and $G1_K = 1_K$. Second, note that $\Pi G(B)$ is uninformative if and only if all of its rows are equal to each other, that is if and only if

$$D(\Pi)G(B) = D(\Pi)(\alpha I_K + (1-\alpha)B) = 0_{(N-1)\times K}, \tag{14}$$

where $D(\Pi)$ is the row-difference matrix of Π:

$$D(\Pi) \triangleq \begin{pmatrix} (\pi_2 - \pi_1)^\top \\ \vdots \\ (\pi_N - \pi_1)^\top \end{pmatrix},$$

and $\pi_i^\top$ is the ith row of Π. Consider the auxiliary finite linear program:

$$\Lambda(\Pi) \triangleq \max_{G \in \mathbb{R}^{K\times K},\, \alpha \in \mathbb{R}} \quad \alpha \tag{15}$$

$$\text{s.t.} \quad G \geq \alpha I_K,\ G1_K = 1_K, \tag{16}$$

$$D(\Pi)G = 0_{(N-1)\times K}. \tag{17}$$

The first lemma identifies the program's value with MVA; the second establishes the rank bounds and the linearly independent case; the third constructs information structures attaining every value in $[1/N, 1/2]$.

Lemma 5. MVA(Π) = $\Lambda(\Pi)$.

Proof. We need to show that there exists a stochastic matrix B such that $\Pi G(B)$ is Blackwell uninformative if and only if $\alpha \leq \Lambda(\Pi)$.

Only if: For any given α, if B is such that $\Pi G(B)$ is Blackwell uninformative, then we showed that $G(B)$ must satisfy conditions (16-17). By the maximization nature of the problem, if $\alpha > \Lambda(\Pi)$, those conditions cannot be satisfied.

If: If $\alpha \leq \Lambda(\Pi)$, then there exists G that satisfies conditions (16-17) (e.g., the argmax). If $\Lambda(\Pi) = 1$, then B can be arbitrary. Otherwise, set $B = (G - \alpha I_K)/(1 - \alpha)$. It is straightforward that the resulting B is a stochastic matrix and by construction $\Pi G(B)$ is uninformative. □

Lemma 6. $\text{MVA}(\Pi) \in [1/R(\Pi), 1/2]$. *If* $R(\Pi) = K$, *then* $\text{MVA}(\Pi) = 1/K$.

Proof. To ease notation, in the proof we omit the dependence of R on Π.

1.) $\text{MVA}(\Pi) \leq 1/2$.

If α and G satisfy (16-17), then $B = (G - \alpha I_K)/(1 - \alpha)$ is a stochastic matrix and

$$D(\Pi)B = -\frac{\alpha}{1-\alpha} D(\Pi). \tag{18}$$

In other words, the rows of $D(\Pi)$ are left eigenvectors of B associated with eigenvalue $-\alpha/(1-\alpha)$. Since B is stochastic, its spectral radius equals 1. Thus, $|-\alpha/(1-\alpha)| \leq 1$ and $\alpha \leq 1/2$. It follows that $\text{MVA}(\Pi) \leq 1/2$.

2.) If $R = K$, then $\text{MVA}(\Pi) = 1/K$.

If $R = K$, then $K \leq N$ and $\text{rank} D(\Pi) = K - 1$. Thus, $\text{rank}\ker D(\Pi) = K - (K-1) = 1$ and, because $D1_K = 1_{N-1} - 1_{N-1} = 0_{N-1}$, $\ker D = \text{span}\{1_K\}$. Thus, for (G, α) to satisfy (17), every column of G must be a multiple of 1_K. But since G is stochastic, it follows that $\sum_{k=1}^{K} G_{kk} = 1$ and $\min_k G_{kk} \leq 1/K$. To further satisfy (16), it must be that $\alpha \leq 1/K$. Thus, $\text{MVA}(\Pi) \leq 1/K$.

At the same time, if $\alpha \leq 1/K$, then (G, α) satisfy (16-17) for $G = 1/K 1_K 1_K^\top$. In this case, G is uninformative, not only ΠG, so the misaligned adviser can make the signal uninformative about his estimate, not only about the state. It follows that $\text{MVA} \geq 1/K$ and, therefore, $\text{MVA}(\Pi) = 1/K$.

3.) $\text{MVA} \geq 1/R$.

Let $\alpha = 1/R$ (recall that $R \geq 2$). Consider the normed space $(\mathbb{R}^K, \|\cdot\|_1)$ and its linear $(R-1)$-dimensional subspace $\mathbb{W}$ spanned by rows of $D(\Pi)$. By the Auerbach basis theorem,

there exist vectors $w_1, \ldots, w_{R-1} \in \mathbb{W}$ and $x_1, \ldots, x_{R-1} \in \mathbb{R}^K$ such that[33]

$$\|w_i\|_1 = 1, \quad \|x_i\|_\infty = 1, \quad w_i^\top x_j = \delta_{ij}, \quad 1 \le i, j \le R-1.$$

Define the corresponding matrices $W \triangleq (w_1, \ldots, w_{R-1})$, $X \triangleq (x_1, \ldots, x_{R-1})$. By construction,

$$W^\top X = I_{R-1}, \tag{19}$$

and by properties of $D(\Pi)$,

$$W^\top 1_K = 0_{R-1}. \tag{20}$$

Define the vector of weights of rows of W, $\overline{w} \in \mathbb{R}^K$, as $\overline{w}_k \triangleq \sum_{i=1}^{R-1} |w_{i,k}|$. Since $\|w_i\|_1 = 1$, we have

$$\sum_{k=1}^{K} \overline{w}_k = \sum_{i=1}^{R-1} \|w_i\|_1 = R-1. \tag{21}$$

We explicitly construct the desired strategy of the misaligned adviser B as:

$$B = \frac{1}{R-1}(1_K \overline{w}^\top - XW^\top). \tag{22}$$

Nonnegativity. For all j, k,

$$B_{jk} = \frac{1}{R-1}\Big(\sum_{i=1}^{R-1} |w_{i,k}| - \sum_{i=1}^{R-1} x_{i,j} w_{i,k}\Big) \ge 0,$$

because $|x_{i,j}| \le \|x_i\|_\infty = 1$.

[33] By the Auerbach theorem, there exist $w_1, \ldots, w_{R-1} \in \mathbb{W}$ and $\phi_1, \ldots, \phi_{R-1} \in \mathbb{W}^*$ such that $\|w_i\|_1 = 1$, $\|\phi_i\|_{\mathbb{W}^*} = 1$, and $\phi_i(w_j) = \delta_{ij}$ (Section II.E, Lemma 11 in Wojtaszczyk (1991); see also Gershkov et al. (2025) for another recent application). By the Hahn–Banach theorem, these ϕ_i, operating on $\mathbb{W}$, can be extended to $\tilde{\phi}_i$, operating on $\mathbb{R}^K$, without a change in their norm. By the duality between spaces l_1 and l_∞, for each i, there exists $x_i \in \mathbb{R}^K$ such that $\tilde{\phi}_i(z) = z^\top x_i$ and $\|x_i\|_\infty = \|\tilde{\phi}_i\| = 1$. Then $w_i^\top x_j = \tilde{\phi}_j(w_i) = \phi_j(w_i) = \delta_{ij}$ for all $1 \le i, j \le R-1$.

Stochasticity. By (20) and (21):

$$B1_K = \frac{1}{R-1}(1_K(\overline{w}^\top 1_K) - X(W^\top 1_K)) = \frac{1}{R-1}(1_K(R-1) - 0_K) = 1_K.$$

Uninformativeness. By (19) and (20):

$$W^\top B = \frac{1}{R-1}((W^\top 1_K)\overline{w}^\top - (W^\top X)W^\top) = \frac{1}{R-1}(0_{R-1} - W^\top) = -\frac{1}{R-1}W^\top.$$

Since by construction columns of W form a basis in the row space of $D(\Pi)$, it follows that

$$D(\Pi)B = -\frac{1}{R-1}D(\Pi).$$

As $\alpha = 1/R$, this corresponds exactly to constraint (17). The result follows. □

Lemma 7. *For any $N \geq 2$ and $\alpha \in [1/N, 1/2]$, there exist K and Π such that* $\mathrm{MVA}(\Pi) = \alpha$.

Proof. The proof is by direct construction. For $N = 2$, the result is trivial. For $N \geq 3$, consider $K \in [4, N+1]$ and, for $\delta \in [0, 1]$, the $N \times K$ matrix Π such that

$$\begin{aligned}
\pi_i^\top &= \frac{1}{K}1_K^\top, \quad i = 1 \text{ or } i = K, K+1, \ldots, N, \\
\pi_i^\top &= \frac{1}{K}(1_K + e_K^i - e_K^1)^\top, \quad i = 2, \ldots, K-2, \\
\pi_i^\top &= \frac{1}{K}(1_K + e_K^i - \delta e_K^1 - (1-\delta)e_K^K)^\top, \quad i = K-1,
\end{aligned}$$

where e_K^i is the ith basis vector of $\mathbb{R}^K$. By construction, Π is a stochastic matrix. Consider $\mathrm{MVA}(\Pi)$ that solves the corresponding problem (15).

The constraint $D(\Pi)G = 0_{(N-1)\times K}$ reduces to:

$$(e_K^i - e_K^1)^\top G = 0, \quad i = 2, \ldots, K-2, \qquad (e_K^{K-1} - \delta e_K^1 - (1-\delta)e_K^K)^\top G = 0, \qquad (23)$$

which effectively states that the first $K-2$ rows are equal to each other and the $(K-1)$th row is a convex combination of the 1st and the Kth rows with weight δ. Thus, the effective variables are the 1st and the Kth rows of the matrix G. The constraints $G \geq \alpha I_{K\times K}$ and

$G1_K = 1_K$ then reduce to those rows being probability vectors, such that

$$G_{1k} \geq \alpha,\ k = 1, \ldots, K-2,\ (\delta G_{1,K-1} + (1-\delta) G_{K,K-1}) \geq \alpha,\ G_{KK} \geq \alpha.$$

Therefore,

$$\alpha \leq (\delta G_{1,K-1} + (1-\delta) G_{K,K-1}) \leq \delta(1-(K-2)\alpha) + (1-\delta)(1-\alpha) = 1 - \alpha\big(1 + \delta(K-3)\big).$$

Rearranging yields

$$\alpha \leq \alpha^\dagger \triangleq \frac{1}{2 + \delta(K-3)}, \tag{24}$$

and thus $\mathrm{MVA}(\Pi) \leq \alpha^\dagger$. Whenever $\delta \geq (K-4)/(K-3)$, $\alpha^\dagger \leq 1/(K-2)$ and the bound $\alpha^\dagger$ can be attained by G with the 1st and the Kth rows being (the rest of G is pinned down by condition (23)):

$$\begin{aligned} &G_{1k} = \alpha^\dagger,\ k = 1, \ldots, K-2, && G_{1,K-1} = 1 - (K-2)\alpha^\dagger, && G_{1K} = 0, \\ &G_{Kk} = 0,\ k = 1, \ldots, K-2, && G_{K,K-1} = 1 - \alpha^\dagger, && G_{KK} = \alpha^\dagger. \end{aligned}$$

Thus, $\mathrm{MVA}(\Pi) = \alpha^\dagger$. At $\delta = (K-4)/(K-3)$, $\alpha^\dagger = 1/(K-2)$; at $\delta = 1$, $\alpha^\dagger = 1/(K-1)$.

This establishes that, for all $K \in [4, N+1]$, as δ spans $[(K-4)/(K-3), 1]$, the proposed Π achieves $\mathrm{MVA}(\Pi)$ that spans $[1/(K-1), 1/(K-2)]$. Spanning K from 4 to $N+1$, we obtain the result. □

Infinite M. We now extend the bounds to arbitrary supports of the adviser's beliefs. Recall the compact set Γ of joint distributions of true beliefs and messages from Lemma 4. For any $\chi \in \Gamma$, the overall message is uninformative about the state if and only if the agent's posterior belief equals the prior μ_0 almost surely. Equivalently, for every continuous function $\varphi : M \to \mathbb{R}$,

$$L_{\alpha,\chi}(\varphi) \triangleq \alpha \int_M (s - \mu_0)\varphi(s)\tau(ds) + (1-\alpha) \int_{M \times M} (s - \mu_0)\varphi(m)\chi(ds, dm) = 0.$$

For each φ, this condition defines a closed subset of $[0,1] \times \Gamma$, since $L_{\alpha,\chi}(\varphi)$ is continuous in (α, χ). By compactness, the largest α for which all these conditions hold is attained. Moreover, if the overall message can be made uninformative at some α, the same is true at any smaller alignment probability.

To establish the upper bound, choose two states ω, ω' with distinct conditional distributions of the adviser's belief, and let $\eta = \pi(\cdot \mid \omega) - \pi(\cdot \mid \omega')$. Uninformativeness requires $\alpha\eta + (1-\alpha)\eta\beta = 0$, where $\eta\beta$ denotes the signed measure defined by $(\eta\beta)(A) = \int_M \beta(A \mid s)\eta(ds)$. Since applying the same reporting kernel to two distributions cannot increase their total variation distance, we have $\|\eta\beta\|_{\mathrm{TV}} \leq \|\eta\|_{\mathrm{TV}}$. As $\eta \neq 0$, it follows that $\alpha \leq 1/2$.

To establish the lower bound, we approximate the adviser's information by finite information structures, to which the previously established bound applies. Partition M into finitely many measurable sets of diameter at most ϵ, and replace each belief s by the conditional mean of its set. Denote this mean by $\bar{s}_\epsilon(s)$. The resulting finite information structure has rank at most $R(\tau)$, so the finite-case result guarantees a reporting strategy that makes the overall message uninformative at $\alpha = 1/R(\tau)$. We can now use this strategy to construct a reporting strategy for the original information structure. Conditional on the adviser's true belief s, first apply the finite reporting strategy to the set containing s, obtaining another set in the partition. Then draw the actual message from τ conditional on the selected set. Let $\chi_\epsilon \in \Gamma$ be the resulting joint distribution of true beliefs and messages. Since the finite reporting strategy makes the selected set uninformative about the state, the only remaining difference comes from truthful reporting of the original belief s rather than its conditional mean $\bar{s}_\epsilon(s)$. Specifically, for every continuous $\varphi : M \to \mathbb{R}$,

$$\begin{aligned} L_{\alpha,\chi_\epsilon}(\varphi) &= \alpha \int_M (s - \bar{s}_\epsilon(s))\varphi(s)\tau(ds), \\ \|L_{\alpha,\chi_\epsilon}(\varphi)\| &\leq \alpha\epsilon\|\varphi\|_\infty. \end{aligned}$$

By compactness of Γ, there is a subsequence of χ_ϵ converging weakly as $\epsilon \to 0$. The bound above implies that its limit satisfies $L_{\alpha,\chi}(\varphi) = 0$ for every continuous φ. Thus, messages can be made uninformative at $\alpha = 1/R(\tau)$, establishing the lower bound.

Finally, note that if the misaligned adviser can make the overall message uninformative,

advice has no value in any decision problem, so $\Delta V = 0$. Conversely, when $\alpha > \text{MVA}(\tau)$, the common minimax reporting strategy β^* from Lemma 4 necessarily induces a nondegenerate distribution of posterior beliefs. In a regular decision problem, such information has strictly positive value: a prior-optimal private strategy cannot also be optimal at a distinct posterior, since it would then be optimal at two distinct interior beliefs along the segment joining them, contradicting regularity. By the saddle-point property, the agent's optimal payoff against β^* equals V^*, so $\Delta V > 0$.

A.5 Proof of Lemma 1

By definition of a TRS, the misaligned adviser with signal realization μ minimizes $U(\hat{\sigma}(\mu'), \mu)$ over μ' in the trust region. Recall that the function $U(\hat{\sigma}(\mu'), \mu)$ is linear in μ, and we assumed that $U(\mu) = \max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$ is strictly convex and twice differentiable in μ. This means that $U(\hat{\sigma}(\mu'), \mu)$ is the value at μ of the hyperplane supporting U at μ'. Under our convention that μ is the probability of state 1, this means that

$$U(\hat{\sigma}(\mu'), \mu) = U(\mu') + U'(\mu')(\mu - \mu').$$

By convexity of U, this function is quasi-concave in μ', and hence for all $\mu' \in [\underline{\mu}, \overline{\mu}]$, $U(\hat{\sigma}(\mu'), \mu) \geq \min\{U(\hat{\sigma}(\underline{\mu}), \mu), U(\hat{\sigma}(\overline{\mu}), \mu)\}$. Thus, the misaligned adviser's strategy takes a threshold form. The threshold $b(\underline{\mu}, \overline{\mu})$ is the intersection point of the supporting lines to U at points $\underline{\mu}$ and $\overline{\mu}$:

$$U(\underline{\mu}) + U'(\underline{\mu})(b(\underline{\mu}, \overline{\mu}) - \underline{\mu}) = U(\overline{\mu}) + U'(\overline{\mu})(b(\underline{\mu}, \overline{\mu}) - \overline{\mu}).$$

Rearranging, we obtain:

$$b(\underline{\mu}, \overline{\mu}) = \frac{\overline{\mu}U'(\overline{\mu}) - \underline{\mu}U'(\underline{\mu}) - (U(\overline{\mu}) - U(\underline{\mu}))}{U'(\overline{\mu}) - U'(\underline{\mu})}.$$

Applying integration by parts, the numerator equals $\int_{\underline{\mu}}^{\overline{\mu}} xU''(x)dx$ and the denominator equals $\int_{\underline{\mu}}^{\overline{\mu}} U''(x)dx$. The result follows.

A.6 Proof of Proposition 1

At $\alpha = 1$, full trust is optimal, so $T = [0, 1]$. When $\alpha = 0$, ignoring the adviser is optimal and is rationalized by the constant reporting strategy $\beta(\mu) = \mu_0$. For the rest of the proof, assume $\alpha \in (0, 1)$. By Lemma 1 and Corollary 3, the choice of an optimal strategy for the agent reduces to optimization over the extreme points $\underline{\mu}$, $\overline{\mu}$ of the trust interval with the corresponding payoff:

$$
\begin{aligned}
&V(\underline{\mu}, \overline{\mu}) \triangleq \\
&\alpha \left(\int_0^{\underline{\mu}} (U(\underline{\mu}) + U'(\underline{\mu})(\mu - \underline{\mu}))\tau(\mu)d\mu + \int_{\underline{\mu}}^{\overline{\mu}} U(\mu)\tau(\mu)d\mu + \int_{\overline{\mu}}^{1} (U(\overline{\mu}) + U'(\overline{\mu})(\mu - \overline{\mu}))\tau(\mu)d\mu \right) \\
&+ (1 - \alpha) \left(\int_0^{b(\underline{\mu},\overline{\mu})} (U(\overline{\mu}) + U'(\overline{\mu})(\mu - \overline{\mu}))\tau(\mu)d\mu + \int_{b(\underline{\mu},\overline{\mu})}^{1} (U(\underline{\mu}) + U'(\underline{\mu})(\mu - \underline{\mu}))\tau(\mu)d\mu \right).
\end{aligned}
$$

The function $V(\underline{\mu}, \overline{\mu})$ is continuously differentiable with partial derivatives (whenever $\underline{\mu} < \overline{\mu}$):

$$
\begin{aligned}
\frac{\partial V}{\partial \underline{\mu}} &= U''(\underline{\mu}) \left(\alpha \int_0^{\underline{\mu}} (\mu - \underline{\mu})\tau(\mu)d\mu + (1 - \alpha) \int_{b(\underline{\mu},\overline{\mu})}^{1} (\mu - \underline{\mu})\tau(\mu)d\mu \right), \\
\frac{\partial V}{\partial \overline{\mu}} &= U''(\overline{\mu}) \left(\alpha \int_{\overline{\mu}}^{1} (\mu - \overline{\mu})\tau(\mu)d\mu + (1 - \alpha) \int_0^{b(\underline{\mu},\overline{\mu})} (\mu - \overline{\mu})\tau(\mu)d\mu \right).
\end{aligned}
$$

Intuitively, the first-order impact of a change in the trust boundary equals the change in the action played at that boundary, measured by $U''(\cdot)$, integrated over the belief regions in which the aligned and misaligned advisers induce that action, weighted by the alignment parameter. (Terms involving $\partial b/\partial \underline{\mu}$ and $\partial b/\partial \overline{\mu}$ vanish because at $\mu = b(\underline{\mu}, \overline{\mu})$ the misaligned adviser is indifferent between the two messages.)

Whenever the trust region is non-singleton, $\underline{\mu} < \overline{\mu}$, at the optimal choice of $\underline{\mu}$ and $\overline{\mu}$ these partial derivatives must equal zero, $\partial V/\partial \underline{\mu} = 0$ and $\partial V/\partial \overline{\mu} = 0$. Since $U''(\cdot) > 0$, these

first-order conditions can be rearranged as follows. Define functions Ψ_1 and Ψ_2 as

$$\Psi_1(\underline{\mu},\overline{\mu},\alpha) \triangleq \alpha\int_0^{\underline{\mu}}(\mu-\underline{\mu})\tau(\mu)d\mu + (1-\alpha)\int_{b(\underline{\mu},\overline{\mu})}^1(\mu-\underline{\mu})\tau(\mu)d\mu,$$
$$\Psi_2(\underline{\mu},\overline{\mu},\alpha) \triangleq \alpha\int_{\overline{\mu}}^1(\mu-\overline{\mu})\tau(\mu)d\mu + (1-\alpha)\int_0^{b(\underline{\mu},\overline{\mu})}(\mu-\overline{\mu})\tau(\mu)d\mu.$$

Then, $\Psi_1(\underline{\mu},\overline{\mu},\alpha)=\Psi_2(\underline{\mu},\overline{\mu},\alpha)=0$ is equivalent to conditions (8) and (9).

First, we show that conditions (8) and (9) are incompatible with $\alpha<1/2$. Indeed, if those conditions hold then (for the rest of the proof, we will often omit the arguments of the function b for brevity):

$$\begin{aligned}\alpha\left(\int_0^b(b-\mu)\tau(\mu)d\mu+\int_b^1(\mu-b)\tau(\mu)d\mu\right) &\geq \alpha\left(\int_0^{\underline{\mu}}(\underline{\mu}-\mu)\tau(\mu)d\mu+\int_{\overline{\mu}}^1(\mu-\overline{\mu})\tau(\mu)d\mu\right)\\ &= (1-\alpha)\left(\int_0^b(\overline{\mu}-\mu)\tau(\mu)d\mu+\int_b^1(\mu-\underline{\mu})\tau(\mu)d\mu\right)\\ &\geq (1-\alpha)\left(\int_0^b(b-\mu)\tau(\mu)d\mu+\int_b^1(\mu-b)\tau(\mu)d\mu\right),\end{aligned}$$

where the inequalities hold because $\underline{\mu}\leq b(\underline{\mu},\overline{\mu})\leq\overline{\mu}$ and the equality is a consequence of (8) and (9). Because τ has full support, the multipliers on both sides of the inequality are strictly positive, and thus $\alpha\geq 1-\alpha$, i.e., $\alpha\geq 1/2$.

Now we argue that for $\alpha\in[1/2,1)$ the solution to (8) and (9) such that $\underline{\mu}\leq\mu_0\leq\overline{\mu}$ exists. Note that at $\alpha=1/2$, $[\underline{\mu},\overline{\mu}]=[\mu_0,\mu_0]$ is a solution.

For the rest of this proof, we omit the dependence of Ψ_i on α. By Lemma 1 and direct inspection, $b(\underline{\mu},\overline{\mu})$ is strictly and continuously increasing in its arguments, so $\Psi_1(\underline{\mu},\overline{\mu})$ is strictly and continuously decreasing in $\underline{\mu}$ for each $\overline{\mu}$. Furthermore, for every $\overline{\mu}\in[\mu_0,1]$,

$$\Psi_1(0,\overline{\mu}) = (1-\alpha)\int_{b(0,\overline{\mu})}^1\mu\tau(\mu)d\mu>0,$$
$$\Psi_1(\mu_0,\overline{\mu}) \leq (1-2\alpha)\int_{\mu_0}^1(\mu-\mu_0)\tau(\mu)d\mu\leq 0,$$

where the second inequality follows from $b(\mu_0,\overline{\mu})\geq\mu_0$ and Bayes plausibility. Therefore, for each $\overline{\mu}\in[\mu_0,1]$, a best-response $b_1(\overline{\mu})\in[0,\mu_0]$ such that $\Psi_1(b_1(\overline{\mu}),\overline{\mu})=0$ exists and is

unique. Since $\Psi_1(\underline{\mu}, \overline{\mu})$ strictly decreases in $\overline{\mu}$, $b_1(\overline{\mu})$ strictly decreases in $\overline{\mu}$.

Analogously, for each $\underline{\mu} \in [0, \mu_0]$, a best-response $b_2(\underline{\mu}) \in [\mu_0, 1]$ such that $\Psi_2(\underline{\mu}, b_2) = 0$ exists, is unique, and strictly decreases in $\underline{\mu}$.

Therefore, a solution to (8) and (9) is any $\underline{\mu} \in [0, \mu_0]$ and $\overline{\mu} = b_2(\underline{\mu}) \in [\mu_0, 1]$ such that $b_1(b_2(\underline{\mu})) = \underline{\mu}$. By the established properties of b_1 and b_2, $b_1(b_2(\underline{\mu}))$ is continuous in $\underline{\mu}$ with $b_1(b_2(\underline{\mu})) \in [0, \mu_0]$ for all $\underline{\mu} \in [0, \mu_0]$; hence, $b_1(b_2(0)) - 0 \geq 0$ and $b_1(b_2(\mu_0)) - \mu_0 \leq 0$. By the intermediate value theorem, there exists $\underline{\mu} \in [0, \mu_0]$ such that $b_1(b_2(\underline{\mu})) = \underline{\mu}$.

So far, we showed that for $\alpha \geq 1/2$, a solution exists and belongs to a closed rectangular set $D = \{(\underline{\mu}, \overline{\mu}) : \underline{\mu} \in [0, \mu_0], \overline{\mu} \in [\mu_0, 1]\}$. To establish uniqueness, consider the function $-\Psi = (-\Psi_1, -\Psi_2)$ on D. Any solution must satisfy $-\Psi(\underline{\mu}, \overline{\mu}) = (0, 0)$. Observe that for any $(\underline{\mu}, \overline{\mu}) \in D$,

$$\begin{aligned}
\frac{\partial[-\Psi_1]}{\partial \underline{\mu}} &= \alpha \int_0^{\underline{\mu}} \tau(\mu) d\mu + (1-\alpha)\frac{\partial b}{\partial \underline{\mu}} \tau(b)(b - \underline{\mu}) + (1-\alpha) \int_b^1 \tau(\mu) d\mu > 0, \\
\frac{\partial[-\Psi_1]}{\partial \overline{\mu}} &= (1-\alpha)\frac{\partial b}{\partial \overline{\mu}} \tau(b)(b - \underline{\mu}) \geq 0, \\
\frac{\partial[-\Psi_2]}{\partial \underline{\mu}} &= -(1-\alpha)\frac{\partial b}{\partial \underline{\mu}} \tau(b)(b - \overline{\mu}) \geq 0, \\
\frac{\partial[-\Psi_2]}{\partial \overline{\mu}} &= \alpha \int_{\overline{\mu}}^1 \tau(\mu) d\mu + (1-\alpha)\frac{\partial b}{\partial \overline{\mu}} \tau(b)(\overline{\mu} - b) + (1-\alpha) \int_0^b \tau(\mu) d\mu > 0.
\end{aligned}$$

Moreover, for all $(\underline{\mu}, \overline{\mu}) \in D$, the Jacobian of $[-\Psi]$ is a P-matrix, i.e., it has strictly positive principal minors:

$$\frac{\partial[-\Psi_1]}{\partial \underline{\mu}} > 0, \quad \frac{\partial[-\Psi_1]}{\partial \underline{\mu}} \frac{\partial[-\Psi_2]}{\partial \overline{\mu}} - \frac{\partial[-\Psi_1]}{\partial \overline{\mu}} \frac{\partial[-\Psi_2]}{\partial \underline{\mu}} > 0.$$

By the Gale-Nikaido Theorem (Gale and Nikaido (1965), Theorem 4), it follows that $[-\Psi]$ is injective on D, and thus there exists at most one solution to the equation $-\Psi(\underline{\mu}, \overline{\mu}) = (0, 0)$.

It remains to establish that the constructed TRS is optimal and robustly rationalizable. By Theorem 1, an optimal TRS exists. For $\alpha \in (1/2, 1)$, Theorem 3 implies that its trust region is nondegenerate. The expressions for the partial derivatives of V show that an optimal interval must also satisfy $0 < \underline{\mu} < \overline{\mu} < 1$. Moreover, the optimal interval contains μ_0, as follows from Theorem 2 and Bayes plausibility. Hence its endpoints must satisfy the

first-order conditions in the domain D. Since these conditions have a unique solution in D, the proposed interval is optimal. Robust rationalizability then follows from Theorem 2.

A.7 Proof of Proposition 2

At $\alpha = 1/2$, $[\underline{\mu}, \overline{\mu}] = [\mu_0, \mu_0]$ satisfies conditions (8) and (9), using the continuous extension of b. At $\alpha = 1$, full trust is optimal, so $[\underline{\mu}, \overline{\mu}] = [0, 1]$.

For $\alpha \in (1/2, 1)$, denote by Ψ_{i1}, Ψ_{i2}, and $\Psi_{i\alpha}$ the partial derivatives of Ψ_i with respect to $\underline{\mu}$, $\overline{\mu}$, and α respectively. Define the Jacobian:

$$J(\underline{\mu}, \overline{\mu}, \alpha) \triangleq \begin{pmatrix} \Psi_{11} & \Psi_{12} \\ \Psi_{21} & \Psi_{22} \end{pmatrix}.$$

As we argued in the proof of Proposition 1, for all $\underline{\mu}$, $\overline{\mu}$, $\alpha > 1/2$, $\det J(\underline{\mu}, \overline{\mu}, \alpha) > 0$, and therefore, by the implicit function theorem, optimal $\underline{\mu}(\alpha)$ and $\overline{\mu}(\alpha)$ are continuously differentiable and[34]

$$\begin{pmatrix} d\underline{\mu}/d\alpha \\ d\overline{\mu}/d\alpha \end{pmatrix} = -J(\underline{\mu}, \overline{\mu}, \alpha)^{-1} \begin{pmatrix} \Psi_{1\alpha} \\ \Psi_{2\alpha} \end{pmatrix}.$$

Consequently,

$$\frac{d\underline{\mu}}{d\alpha} = -\frac{\Psi_{2\overline{\mu}}\Psi_{1\alpha} - \Psi_{1\overline{\mu}}\Psi_{2\alpha}}{\Psi_{1\underline{\mu}}\Psi_{2\overline{\mu}} - \Psi_{1\overline{\mu}}\Psi_{2\underline{\mu}}} < 0,$$
$$\frac{d\overline{\mu}}{d\alpha} = \frac{\Psi_{2\underline{\mu}}\Psi_{1\alpha} - \Psi_{1\underline{\mu}}\Psi_{2\alpha}}{\Psi_{1\underline{\mu}}\Psi_{2\overline{\mu}} - \Psi_{1\overline{\mu}}\Psi_{2\underline{\mu}}} > 0,$$

where the inequalities hold because, as we already showed, $\Psi_{1\underline{\mu}} < 0$, $\Psi_{1\overline{\mu}} \leq 0$, $\Psi_{2\underline{\mu}} \leq 0$,

[34] Differentiating the optimality conditions with respect to α we obtain $\Psi_{11}\frac{d\underline{\mu}}{d\alpha} + \Psi_{12}\frac{d\overline{\mu}}{d\alpha} + \Psi_{1\alpha} = 0, \Psi_{21}\frac{d\underline{\mu}}{d\alpha} + \Psi_{22}\frac{d\overline{\mu}}{d\alpha} + \Psi_{2\alpha} = 0$.

$\Psi_{2\overline{\mu}} < 0$, and

$$\Psi_{1\alpha} = \int_0^{\underline{\mu}} (\mu - \underline{\mu})\tau(\mu)d\mu - \int_b^1 (\mu - \underline{\mu})\tau(\mu)d\mu < 0,$$
$$\Psi_{2\alpha} = \int_{\overline{\mu}}^1 (\mu - \overline{\mu})\tau(\mu)d\mu - \int_0^b (\mu - \overline{\mu})\tau(\mu)d\mu > 0.$$

The result follows.

A.8 Proof of Proposition 3

Throughout, fix $\alpha \in (1/2, 1)$; the cases $\alpha \leq 1/2$ and $\alpha = 1$ are immediate. We begin with a simple lemma.

Lemma 8. *Let U_1, U_2 be twice differentiable and strictly convex on $[0,1]$. Assume that the ratio $U_1''(\mu)/U_2''(\mu)$ is decreasing. Then, with b_i defined for each U_i, $i \in \{1, 2\}$, by equation (7), we have that $b_1(\underline{\mu}, \overline{\mu}) \leq b_2(\underline{\mu}, \overline{\mu})$ for all $\underline{\mu} \leq \overline{\mu}$.*

Proof. By equation (7), for $i \in \{1, 2\}$, $b_i(\underline{\mu}, \overline{\mu}) = \mathbb{E}_{\mu \sim f_i}[\mu]$, where $f_i(\mu)$ is a density of a probability measure on $[\underline{\mu}, \overline{\mu}]$ defined as

$$f_i(\mu) \triangleq \frac{U_i''(\mu)}{\int_{\underline{\mu}}^{\overline{\mu}} U_i''(\mu)\, d\mu}.$$

By assumption, $f_1(\mu)/f_2(\mu)$ is decreasing, and hence f_2 likelihood-ratio dominates f_1. This implies that the probability distribution f_1 is first-order stochastically dominated by f_2; in particular, it has a lower mean. The result follows. □

We now take the second step by showing a monotone relationship between the cutoff function b and the trust region.

Lemma 9. *Consider two decision problems U_i, $i \in \{1, 2\}$, and let $(\underline{\mu}_i, \overline{\mu}_i) \in D$ denote the unique solution to the system*

$$\Psi_1^i(\underline{\mu}, \overline{\mu}, \alpha) = 0, \qquad \Psi_2^i(\underline{\mu}, \overline{\mu}, \alpha) = 0,$$

where Ψ_1^i, Ψ_2^i *and* D *are defined as in the proof of Proposition 1 for each* U_i*. If* $b_1(\underline{\mu}, \overline{\mu}) \leq b_2(\underline{\mu}, \overline{\mu})$ *for all* $\underline{\mu} \leq \overline{\mu}$*, then* $\underline{\mu}_2 \leq \underline{\mu}_1$ *and* $\overline{\mu}_2 \leq \overline{\mu}_1$*.*

Proof. For any $h \in [0,1]$, define the auxiliary functions

$$\widetilde{\Psi}_1(\underline{\mu}, h) \triangleq \alpha \int_0^{\underline{\mu}} (\mu - \underline{\mu})\tau(\mu)\, d\mu + (1-\alpha) \int_h^1 (\mu - \underline{\mu})\tau(\mu)\, d\mu,$$
$$\widetilde{\Psi}_2(\overline{\mu}, h) \triangleq \alpha \int_{\overline{\mu}}^1 (\mu - \overline{\mu})\tau(\mu)\, d\mu + (1-\alpha) \int_0^h (\mu - \overline{\mu})\tau(\mu)\, d\mu.$$

For each h, let $\underline{\mu}(h)$ be the unique solution to $\widetilde{\Psi}_1(\underline{\mu}, h) = 0$ in $[0, \mu_0]$, and let $\overline{\mu}(h)$ be the unique solution to $\widetilde{\Psi}_2(\overline{\mu}, h) = 0$ in $[\mu_0, 1]$, with the existence and uniqueness following from an argument analogous to that used in the proof of Proposition 1.

For each $i \in \{1, 2\}$ define the scalar mapping

$$\varphi_i(h) \;\triangleq\; b_i\big(\underline{\mu}(h), \overline{\mu}(h)\big).$$

Let $h_i \triangleq b_i(\underline{\mu}_i, \overline{\mu}_i)$ be the cutoff evaluated at the optimal endpoints of problem i. Then $(\underline{\mu}_i, \overline{\mu}_i)$ solves $\Psi_1^i = \Psi_2^i = 0$ if and only if $\underline{\mu}_i = \underline{\mu}(h_i)$, $\overline{\mu}_i = \overline{\mu}(h_i)$, and $h_i = \varphi_i(h_i)$. By the assumption $b_1 \leq b_2$ pointwise, for every h,

$$\varphi_1(h) = b_1(\underline{\mu}(h), \overline{\mu}(h)) \;\leq\; b_2(\underline{\mu}(h), \overline{\mu}(h)) = \varphi_2(h).$$

Define a region $H \triangleq \{h \in [0,1] : \underline{\mu}(h) \leq h \leq \overline{\mu}(h)\}$. Note that H is an interval and $h_1, h_2 \in H$. For $h \in H$, implicit differentiation yields

$$\underline{\mu}'(h) = -\frac{\partial \widetilde{\Psi}_1 / \partial h}{\partial \widetilde{\Psi}_1 / \partial \underline{\mu}} \leq 0, \qquad \overline{\mu}'(h) = -\frac{\partial \widetilde{\Psi}_2 / \partial h}{\partial \widetilde{\Psi}_2 / \partial \overline{\mu}} \leq 0.$$

Thus, φ_i is weakly decreasing in h on H: both $\underline{\mu}(h)$ and $\overline{\mu}(h)$ are weakly decreasing in h, while $b_i(\underline{\mu}, \overline{\mu})$ is weakly increasing in each endpoint, so the composition $h \mapsto \varphi_i(h)$ is weakly decreasing.

It follows that $h_2 \geq h_1$: if $h_2 < h_1$, then, since φ_2 is decreasing on H,

$$h_2 = \varphi_2(h_2) \geq \varphi_2(h_1) \geq \varphi_1(h_1) = h_1,$$

which is a contradiction. Since $\underline{\mu}(\cdot)$ and $\overline{\mu}(\cdot)$ are weakly decreasing,

$$\underline{\mu}_2 = \underline{\mu}(h_2) \leq \underline{\mu}(h_1) = \underline{\mu}_1, \qquad \overline{\mu}_2 = \overline{\mu}(h_2) \leq \overline{\mu}(h_1) = \overline{\mu}_1,$$

completing the proof. □

By Lemma 8, U_1''/U_2'' decreasing implies $b_1(\underline{\mu}, \overline{\mu}) \leq b_2(\underline{\mu}, \overline{\mu})$ for all $\underline{\mu} \leq \overline{\mu}$. Lemma 9 then yields $\underline{\mu}_2 \leq \underline{\mu}_1$ and $\overline{\mu}_2 \leq \overline{\mu}_1$. This proves Proposition 3.

B Additional results

B.1 Sufficient Conditions for Regularity

In this appendix, we show that the agent's decision problem is regular when her private information induces a full-support distribution of beliefs with a density, under additional genericity assumptions.

We assume that the action space is finite and the agent's ex-post payoff is type-independent, $u(a, \omega)$. It is convenient to identify the agent's type θ with the posterior belief it induces under the reference prior μ_0. Thus, $\Theta = \Delta(\Omega)$ and $\theta(\omega)$ is the posterior probability of state ω. For an arbitrary initial belief μ, the agent's posterior after observing θ is

$$\nu_{\mu,\theta}(\omega) \triangleq \frac{\frac{\mu(\omega)}{\mu_0(\omega)}\theta(\omega)}{\sum_{\omega' \in \Omega} \frac{\mu(\omega')}{\mu_0(\omega')}\theta(\omega')}. \tag{25}$$

A necessary and sufficient condition for a private strategy $\hat{\sigma}$ to be Bayes-optimal at belief μ is that, for almost every θ under the type distribution induced by μ, $\hat{\sigma}(\cdot|\theta) \in \Delta(A)$ is an

optimal best-response with respect to $\nu_{\mu,\theta}$: for all $a \in \operatorname{supp} \hat{\sigma}(\cdot|\theta)$,

$$a \in \arg\max_{a' \in A} \sum_{\omega \in \Omega} \nu_{\mu,\theta}(\omega) u(a', \omega). \tag{26}$$

Assumption 1. Any two distinct actions yield different payoffs in every state: $u(a, \omega) \neq u(a', \omega)$ for all $a \neq a'$ and $\omega \in \Omega$. Moreover, there exist two distinct actions $a_1, a_2 \in A$ and an interior belief $\mu \in \Delta(\Omega)$ such that $\mathbb{E}_\mu[u(a_1, \omega)] = \mathbb{E}_\mu[u(a_2, \omega)] > \mathbb{E}_\mu[u(a, \omega)]$, for all $a \notin \{a_1, a_2\}$.

The first condition is restrictive but holds generically for finite state and action spaces. It ensures that, for any belief, the agent's optimal action is unique for almost every private type. The second condition requires that exactly two actions be optimal at some interior belief. Together with the first condition, it ensures that any change in the agent's prior shifts the boundary between private types who prefer one action over the other. With sufficiently rich private information, this implies that any change in beliefs is decision-relevant for some agent types.

Lemma 10. *Suppose θ has a full-support distribution on $\Delta(\Omega)$ with a density with respect to Lebesgue measure under the reference prior μ_0, and Assumption 1 holds. Then the agent's decision problem is regular.*

Proof. We first establish part (i) of regularity. Let $\bar{f}$ denote the distribution of θ under μ_0. Since θ is the agent's posterior under μ_0, Bayes' rule implies

$$f(d\theta \mid \omega) = \frac{\theta(\omega)}{\mu_0(\omega)} \bar{f}(d\theta).$$

Since θ belongs to the interior of the simplex almost surely, the distributions of types under μ and under μ_0 are mutually absolutely continuous, for any belief μ. Thus, when referring to events that occur with zero or full probability, we need not specify the belief under which probabilities are evaluated.

Fix $\mu \in \Delta(\Omega)$ and two distinct actions $a, a' \in A$. These actions can be simultaneously

optimal only if

$$\sum_{\omega\in\Omega}\frac{\mu(\omega)}{\mu_0(\omega)}\,\theta(\omega)\big[u(a,\omega)-u(a',\omega)\big]=0.$$

By Assumption 1, the actions have different payoffs in every state. Thus, this equality defines a proper hyperplane in the belief simplex (or an empty set), which has measure zero since θ has a density. As A is finite, the optimal action is therefore unique for almost every type, under any belief μ. Since the state-conditional distributions of types are absolutely continuous with respect to $\bar{f}$, all optimal decision rules induce the same profile of state-contingent expected payoffs.

We next establish part (ii). Fix any distinct interior beliefs μ_1,μ_2, and let a_1,a_2 and μ^* be as in Assumption 1. Define

$$d(\omega)\triangleq u(a_1,\omega)-u(a_2,\omega),\qquad L_i(\theta)\triangleq\sum_{\omega\in\Omega}\frac{\mu_i(\omega)}{\mu_0(\omega)}\theta(\omega)d(\omega),\quad i=1,2.$$

There exists an interior type θ^* whose posterior under μ_1 is μ^*. Hence $L_1(\theta^*)=0$, and a_1,a_2 strictly outperform all other actions at θ^* under μ_1. By continuity, the latter property continues to hold in some open neighborhood O of θ^*.

Since $\mu_1\neq\mu_2$ and $d(\omega)\neq 0$ for every ω, the hyperplanes $\{\theta: L_1(\theta)=0\}$ and $\{\theta: L_2(\theta)=0\}$ are distinct. Consequently, there exists $\bar{\theta}\in O$ such that $L_1(\bar{\theta})=0$ and $L_2(\bar{\theta})\neq 0$. Without loss of generality, suppose $L_2(\bar{\theta})<0$; otherwise interchange a_1 and a_2. By continuity, there exists a nonempty open set $B\subset O$ such that

$$L_1(\theta)>0\quad\text{and}\quad L_2(\theta)<0\qquad\text{for every }\theta\in B.$$

For every $\theta\in B$, action a_1 is uniquely optimal under μ_1, but is not optimal under μ_2. Since B has positive probability under both beliefs, no decision rule can be Bayes-optimal at both μ_1 and μ_2. This proves part (ii). □

B.2 Analytical solutions with multiple states

In this appendix, we present solutions to two tractable cases within the multi-state environment. First, we provide a complete characterization for the case of binary actions and no private information of the agent. Then, we present a solution to a parametric class of problems in which symmetry permits an explicit characterization of the optimal trust region.

B.2.1 Binary Action

Consider the setting in which the agent has only two pure private strategies, i.e., $A = \{a_1, a_2\}$ and $|\Theta| = 1$. Thus, the agent has no private information and we drop the type throughout.

Without loss of generality, we can normalize the agent's payoff from action a_1 to zero, $u(a_1, \omega) \equiv 0$, and denote the expected payoff from action a_2 when the adviser's posterior is μ by $v(\mu) \triangleq \mathbb{E}_\mu[u(a_2, \omega)]$. Denote by $\hat{\tau} \in \Delta(\mathbb{R})$ the distribution of v when μ is distributed according to $\tau \in \Delta(\Delta(\Omega))$.

Define the absolute losses and gains from taking the second action relative to the first one:

$$L(\hat{\tau}) = \int_{-\infty}^{0} (-v)\, \hat{\tau}(dv), \quad G(\hat{\tau}) = \int_{0}^{+\infty} v\, \hat{\tau}(dv). \tag{27}$$

Also, define the following threshold:

$$\hat{\alpha}(\hat{\tau}) = \frac{\max\{L(\hat{\tau}), G(\hat{\tau})\}}{L(\hat{\tau}) + G(\hat{\tau})}. \tag{28}$$

To rule out trivial cases and to simplify the exposition of the optimal strategy, we make the following assumption that holds in generic non-degenerate environments.

Assumption 2 (Generic Nondegeneracy). $L(\hat{\tau}) > 0$, $G(\hat{\tau}) > 0$, $L(\hat{\tau}) \neq G(\hat{\tau})$, $\tau(\{\mu : v(\mu) = 0\}) = 0$.

Proposition 4 (Binary Action Solution). *Suppose that Assumption 2 holds. If $\alpha \neq \hat{\alpha}(\hat{\tau})$, then the optimal solution exists, is unique, and is robustly rationalizable. The trust region can be chosen as follows: if $\alpha > \hat{\alpha}(\hat{\tau})$, then all messages are trusted, $T = \Delta(\Omega)$; if $\alpha < \hat{\alpha}(\hat{\tau})$, then no messages are trusted, $T = \{\mu_0\}$. If $\alpha = \hat{\alpha}(\hat{\tau})$, both full trust and no trust are optimal and robustly rationalizable.*

By Proposition 4, generically, the optimal solution is stark: either all or none of the adviser's messages are trusted. This is in contrast to the binary-state case with a rich strategy space, where the trust region expanded continuously with the alignment probability (Proposition 2).

Notably, only the aggregate quantities $L(\hat{\tau})$ and $G(\hat{\tau})$ matter for the determination of the trust region; the detailed distribution of relative payoffs $\hat{\tau}$ is irrelevant. The threshold $\hat{\alpha}(\hat{\tau})$ is minimized at $L(\hat{\tau}) = G(\hat{\tau})$, in which case $\hat{\alpha}(\hat{\tau}) = 1/2$. Hence, if $\alpha < 1/2$, the agent never trusts the adviser, regardless of $\hat{\tau}$.[35] By contrast, for a given $\alpha > 1/2$, the trust condition $\hat{\alpha}(\hat{\tau}) \leq \alpha$ is equivalent to $(L(\hat{\tau}), G(\hat{\tau}))$ lying in the cone defined by the two linear inequalities

$$(1-\alpha)L(\hat{\tau}) \leq \alpha G(\hat{\tau}), \quad (1-\alpha)G(\hat{\tau}) \leq \alpha L(\hat{\tau}).$$

Thus, in binary decision problems, the adviser is beneficial only when the expected gains and losses of one action relative to the other are not too far apart.

Proof of Proposition 4. With a small abuse of notation, we can parameterize each private strategy by $\hat{\sigma} = \Pr(a = a_2)$. We also drop the dependence of G and L on $\hat{\tau}$ in the notation. Then, by the arguments behind Theorem 1, if the agent employs the set of private strategies $\hat{\Sigma}_0 = \{\hat{\sigma}(m)\}_{m \in \Delta(\Omega)}$, the payoffs coming from both aligned and misaligned adviser depend only on $\hat{\sigma}_L \triangleq \inf \hat{\Sigma}_0$ and $\hat{\sigma}_H \triangleq \sup \hat{\Sigma}_0$, and the optimal payoffs from using $\hat{\Sigma}_0$ are the same as if the agent plays $\hat{\sigma}(m) = \hat{\sigma}_L$ when $v(m) < 0$ and $\hat{\sigma}(m) = \hat{\sigma}_H$ when $v(m) \geq 0$. This payoff is:

$$\begin{aligned} &\int_{-\infty}^{0} (\alpha\hat{\sigma}_L + (1-\alpha)\hat{\sigma}_H) v \, \hat{\tau}(dv) + \int_{0}^{+\infty} (\alpha\hat{\sigma}_H + (1-\alpha)\hat{\sigma}_L) v \, \hat{\tau}(dv) \\ &= \hat{\sigma}_L((1-\alpha)G - \alpha L) + \hat{\sigma}_H(\alpha G - (1-\alpha)L). \end{aligned} \tag{29}$$

The optimal choice of $\hat{\sigma}_L$ and $\hat{\sigma}_H$ must maximize (29) subject to $\hat{\sigma}_L, \hat{\sigma}_H \in [0,1]$ and $\hat{\sigma}_L \leq \hat{\sigma}_H$. This is a linear optimization subject to $(\hat{\sigma}_L, \hat{\sigma}_H)$ being in a triangle with vertices $(0,0)$, $(0,1)$, and $(1,1)$. A straightforward calculation gives the following solution:

[35] This conclusion is driven by the coarseness of the strategy space; we know from Theorem 3 that, for any $\alpha > 0$, the agent would optimally use a non-trivial trust region if the state space and her action space were sufficiently rich.

If $G = L$: if $\alpha < \hat{\alpha} \triangleq 1/2$, then any $\hat{\sigma}_L = \hat{\sigma}_H$ is optimal; if $\alpha > \hat{\alpha}$, then $\hat{\sigma}_L = 0$ and $\hat{\sigma}_H = 1$; if $\alpha = \hat{\alpha}$, then any $(\hat{\sigma}_L, \hat{\sigma}_H)$ is optimal. If $G > L$: if $\alpha < \hat{\alpha}$, then $\hat{\sigma}_L = \hat{\sigma}_H = 1$; if $\alpha > \hat{\alpha}$, then $\hat{\sigma}_L = 0$ and $\hat{\sigma}_H = 1$; if $\alpha = \hat{\alpha}$, then $\hat{\sigma}_H = 1$ and any $\hat{\sigma}_L$ is optimal. If $G < L$: if $\alpha < \hat{\alpha}$, then $\hat{\sigma}_L = \hat{\sigma}_H = 0$; if $\alpha > \hat{\alpha}$, then $\hat{\sigma}_L = 0$ and $\hat{\sigma}_H = 1$; if $\alpha = \hat{\alpha}$, then $\hat{\sigma}_L = 0$ and any $\hat{\sigma}_H$ is optimal.

The cases $\hat{\sigma}_L = \hat{\sigma}_H$ correspond to not trusting any message and always acting in the same way, optimal at the prior belief, so $T = \{\mu_0\}$. The cases $\hat{\sigma}_L = 0$ and $\hat{\sigma}_H = 1$ correspond to trusting all messages, so $T = \Delta(\Omega)$. Since we assumed that the probability of $v(\mu) = 0$ is 0 and $G \neq L$, the corresponding optimal strategy is uniquely determined.

It is left to show that those strategies are robustly rationalizable. Define $M_0 = \{\mu : v(\mu) = 0\}$, $M_- = \{\mu : v(\mu) < 0\}$, and $M_+ = \{\mu : v(\mu) > 0\}$. Define probability measures $q_+(X) = \int_X v(\mu)\tau(d\mu)/G$ for $X \subseteq M_+$, $q_-(Y) = \int_Y (-v(\mu))\tau(d\mu)/L$ for $Y \subseteq M_-$.

For $\alpha > \hat{\alpha}$, the agent fully trusts the adviser. Consider the following strategy of the misaligned adviser. If $\mu \in M_0$, then $\beta(\mu) = \mu$. If $\mu \in M_-$, then β randomizes over messages $m \in M_+$ according to q_+. If $\mu \in M_+$, then β randomizes over messages $m \in M_-$ according to q_-. This strategy is clearly adversarial. Furthermore, since $\alpha > \hat{\alpha}$, after any message $m \in M_+$, the posterior expected payoff from action a_2 is strictly positive: for any $X \subseteq M_+$ with $\tau(X) > 0$,

$$\alpha \int_X v(m)\tau(dm) + (1-\alpha) \int_{\Delta(\Omega)} v(\mu)\beta(X|\mu)\tau(d\mu) = \alpha G q_+(X) - (1-\alpha) L q_+(X) > 0.$$

Analogously, after any message $m \in M_-$, the posterior expected payoff from action a_2 is strictly negative: for any $Y \subseteq M_-$ with $\tau(Y) > 0$,

$$\alpha \int_Y v(m)\tau(dm) + (1-\alpha) \int_{\Delta(\Omega)} v(\mu)\beta(Y|\mu)\tau(d\mu) = -\alpha L q_-(Y) + (1-\alpha) G q_-(Y) < 0.$$

And by construction, after any message $m \in M_0$, the posterior expected payoff from action a_2 is zero.

For $\alpha < \hat{\alpha}$, the agent doesn't trust the adviser so any adviser's strategy is adversarial. Consider the case $G > L$; the complementary case is analogous. We need to construct the misaligned adviser strategy that makes communication not valuable. To do so, define

$\gamma = \alpha L/((1-\alpha)G) \in [0,1]$. Consider the following strategy of the misaligned adviser. If $\mu \in M_-$, then β randomizes over messages $m \in M_+$ according to q_+. If $\mu \in M_+$, then with probability γ, β randomizes over messages $m \in M_-$ according to q_- and with probability $(1-\gamma)$, β randomizes over messages $m \in M_+$ according to q_+. This strategy makes the posterior expected payoff from action a_2 zero after every message $m \in M_-$: for any $Y \subseteq M_-$ with $\tau(Y) > 0$,

$$\alpha \int_Y v(m)\tau(dm) + (1-\alpha) \int_{\Delta(\Omega)} v(\mu)\beta(Y|\mu)\tau(d\mu) = -\alpha L q_-(Y) + (1-\alpha)\gamma G q_-(Y) = 0.$$

After any message $m \in M_+$, the posterior expected payoff from action a_2 is strictly positive: for any $X \subseteq M_+$ with $\tau(X) > 0$,

$$\alpha \int_X v(m)\tau(dm) + (1-\alpha) \int_{\Delta(\Omega)} v(\mu)\beta(X|\mu)\tau(d\mu) = (G-L)q_+(X) > 0.$$

Thus, this β robustly rationalizes the strategy.

Finally, at $\alpha = \hat{\alpha}$, both full trust and no trust are optimal, along with a continuum of other strategies, and robustly rationalizable by the same strategies β as above.

B.2.2 Spherical Environments

In this appendix, we characterize the optimal trust region in radially symmetric environments.

Use coordinates centered at the prior μ_0, so beliefs lie in the translated simplex $\Delta(\Omega) - \mu_0$ and the origin represents the prior. With a slight abuse of notation, write μ for the centered belief and U for the translated indirect utility. Let $U(\mu) = \tilde{U}(\|\mu\|)$ be twice continuously differentiable, with $\tilde{U}'' > 0$. Let the adviser's belief be symmetrically distributed over a ball $C = \{\mu : \|\mu\| \leq r_0\}$ in the affine hull of the translated simplex, contained in $\Delta(\Omega) - \mu_0$, with radial density $\tau(r)$. Then, there exists a robustly rationalizable solution in which the trust region is the ball $T = \{\mu : \|\mu\| \leq r^*(\alpha)\}$.

We will show this result via Theorem 2 by explicitly constructing the corresponding TRE. The key observation, that we formalize and prove in Lemma 11, is that the misaligned adviser with belief μ induces an antipodal belief on the boundary of T. This fact combined with the

symmetry of the problem implies that the adversarial strategy is the same on each ray going through the center of the ball, and hence analogous to the strategy constructed in Section 4.

The radius $r^*(\alpha)$ can be found via the balancing condition applied to any ray going through the center. Indeed, consider any line through the origin, with signed coordinate r and boundary points $-r_0$ and r_0. Then, the belief at $r = r^*(\alpha)$ will be induced by the misaligned adviser only when his belief is at negative coordinates, and by the aligned adviser only when his belief is at positive coordinates. Thus, the agent's interim beliefs satisfy the TRE property if and only if:

$$r^* = \frac{\alpha \int_{r^*}^{r_0} r\tau(r)dr - (1-\alpha)\int_0^{r_0} r\tau(r)dr}{\alpha \int_{r^*}^{r_0} \tau(r)dr + (1-\alpha)\int_0^{r_0} \tau(r)dr}.$$

Rearranging, we obtain:

$$(2\alpha - 1)\int_{r^*}^{r_0} (r - r^*)\tau(r)dr = (1-\alpha)\left(\int_0^{r^*} (r + r^*)\tau(r)dr + \int_{r^*}^{r_0} 2r^*\tau(r)dr\right). \quad (30)$$

For $\alpha < 1/2$, the equation does not admit a solution. In this case, the optimal radius is $r^*(\alpha) = 0$, since the misaligned adviser can make communication uninformative by reporting the antipodal belief with probability $\alpha/(1-\alpha)$ and an independent draw from τ otherwise. At $\alpha = 1/2$, $r^* = 0$ is a solution. For $\alpha \in (1/2, 1)$, the left-hand side is continuously and strictly decreasing in r^*, and the right-hand side is continuously and strictly increasing in r^*, with a derivative with respect to r^* that is strictly positive. Therefore, the equation admits a unique solution $r^*(\alpha)$. As the left-hand side strictly increases in α and the right-hand side strictly decreases in α, $r^*(\alpha)$ strictly increases in α. Furthermore, $r^*(1) = r_0$.

This example features two notable properties. First, the MVA does not depend on U or the number of states; the threshold alignment probability is always $1/2$. Second, the shape of $\tilde{U}$, which captures the details of the decision problem, does not matter for the trust region T; the trust region is uniquely pinned down by $\tau(r)$. For example, if τ is uniform, $\tau \sim U[0, r_0]$, then, for $\alpha \geq 1/2$, $r^*(\alpha) = \frac{1-\sqrt{1+\alpha-2\alpha^2}}{\alpha} r_0$.

Supporting calculations. The key step in the construction of the optimal trust region comes from the following simple lemma.

Lemma 11 (Spherical U). *Let $U(\mu) = \tilde{U}(\|\mu\|)$ be as above and $\mu'(r, n) = rn$, where n is a unit direction parallel to the belief simplex and $r \in [0, r_0]$. Then, for any $\mu \in C$, (i) $D_U(\mu, \mu'(r, n))$ is strictly increasing in $-n \cdot \mu$ whenever $r > 0$, and (ii) $D_U(\mu, \mu'(r, n))$ is a unimodal function in r with a minimum at $r = \max\{0, n \cdot \mu\}$.*

Proof. In this case, $\nabla U(\mu'(r, n)) = \tilde{U}'(r)n$, and thus

$$D_U(\mu, \mu'(r, n)) = \tilde{U}(\|\mu\|) - \tilde{U}(r) - \tilde{U}'(r)(n \cdot \mu - r).$$

By our convexity assumptions on U, $\tilde{U}'(r) > 0$ whenever $r > 0$ and $\tilde{U}''(r) > 0$. Thus $D_U(\mu, \mu'(r, n))$ is strictly increasing in $-n \cdot \mu$ for $r > 0$. Furthermore,

$$\frac{\partial D_U(\mu, \mu'(r, n))}{\partial r} = \tilde{U}''(r)(r - n \cdot \mu).$$

Because $\tilde{U}''(r) > 0$, the unimodality follows. □

Suppose that the misaligned adviser holds belief μ. The adviser will try to maximize $D_U(\mu, \mu')$ over μ' in the trust region. For $\mu \neq 0$, parameterizing $\mu' = rn$ and applying Lemma 11 yields that n is optimally chosen to be $-\mu/\|\mu\|$, whereas r is chosen to be maximal within the trust region, $r = r^*(\alpha)$. Thus, when $r^*(\alpha) > 0$, the optimal μ' is uniquely given by $-r^*(\alpha)\mu/\|\mu\|$. If $\mu = 0$ and $r^*(\alpha) > 0$, every point on the boundary is optimal. If $r^*(\alpha) = 0$, the unique optimal belief is $\mu' = 0$.