ROBUST TRUST*

Piotr Dworczak Ⓡ Alex Smolin

ABSTRACT

We study an agent who combines her private information with recommendations from an informed but potentially misaligned adviser. The adviser observes a signal and, with known probability, reveals it truthfully; otherwise he can send an arbitrary message. We characterize the agent's inference-and-action rule that delivers the maximal guaranteed payoff. Any optimal rule admits a trust region representation in belief space: advice is taken at face value when it induces a posterior within the trust region, and otherwise the agent acts as if the posterior were on the trust region's boundary. We show that commitment has no value to the agent and derive thresholds on the truthfulness probability above which the adviser's presence strictly benefits the agent.

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. 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 ensure safe adoption of AI-aided decision making.

In this paper, we study how an agent 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 agent's optimal robust inference-and-action rule, which maximizes her payoff under the assumption of worst-case AI behavior in case of system misalignment.

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 expertise or contextual information—but she can additionally rely on reports from an adviser (e.g., an AI system). The adviser observes complementary information about the state and sends a message to the agent. Crucially, the adviser is aligned and reports its information truthfully only with some known alignment probability; with remaining probability, she can send an arbitrary message. Facing potential misalignment with the adviser, the agent adopts a robust approach: she chooses a policy mapping her private information and the adviser's message into a (possibly random) action that maximizes her expected payoff against the misaligned adviser who is attempting to minimize her payoff. This criterion reflects the safety concerns in applications: an optimal robust inference-and-action rule provides the highest payoff guarantee for the agent in the absence of any assumptions about the degree and form of misalignment.

Our main structural result shows that optimal robust policy for the agent 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 is willing to take 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 by the notion of Bregman distance) which is a belief lying on the boundary of the trust region; she then behaves as if that belief had been reported.

*Dworczak: Department of Economics, Northwestern University and Group for Research in Applied Economics, piotr.dworczak@northwestern.edu. Smolin: Toulouse School of Economics, alexey.v.smolin@gmail.com. The full version of the paper is at https://arxiv.org/pdf/2602.09490.

Operationally, this is an endogenous form of clipping: moderate recommendations are followed while extreme recommendations are discounted and converted into boundary recommendations that the agent is still willing to accept.

Intuitively, if the agent reacted sharply to extreme reports, the misaligned adviser could exploit that sensitivity to induce large losses. The robust policy responds by limiting how far any recommendation can push behavior. The trust region identifies exactly which recommendations are safe to act upon without additional skepticism, and the boundary mapping 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 reports; on the other extreme, a trust region only containing the prior corresponds to ignoring the adviser's reports. Thus, the shape and size of the trust region yields a disciplined answer to a practical design question: when an AI system outputs highly confident or highly unusual recommendations, optimal robust use requires treating those outputs as "too informative to be trusted" and translating them into safer boundary inputs before acting.

A noteworthy implication of the characterization is that the optimal robust action rule used by the agent must be defensible as optimal for some coherent set of beliefs about the state of the world. In other words, the agent never benefits from distorted use of her own private information; optimal policies only use responses that lie on the payoff frontier of Bayesian decision-making. An optimal robust 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. There exists an equilibrium in which the agent's policy and the misaligned adviser's worst-case strategy form a saddle point of the resulting zero-sum game. In this equilibrium, after every on-path recommendation, the agent's chosen action is Bayes-optimal given the belief induced by the adviser's strategy. 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. Practically, this 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.

Having characterized the structure of optimal decision-making under misalignment concerns, we then ask when consulting a potentially misaligned adviser is worthwhile for the agent. A cautious agent may gain nothing if misalignment is too likely—the trust region could collapse to the prior. We formalize this question by defining "minimal viable alignment"—the threshold alignment probability above which the agent can guarantee a strictly higher payoff than she can achieve using only her own information. Below this threshold, any potential informational benefit from truthful advice can be neutralized by adversarial distortions, and robust value collapses to zero. We derive sharp bounds on this threshold that depend only on the richness of the state space and the adviser's informativeness.

Our general characterization becomes particularly sharp when the state space or strategy space is binary. When the state space is binary—corresponding to situations in which the ground truth is whether a given statement is true or false—an adviser's message can be summarized by the implied probability of the statement being true. The trust region is simply an interval around the prior probability. Recommendations inside the interval are trusted and acted upon as reported. Recommendations outside the interval are mapped to the nearest endpoint. The misaligned adviser sends messages that push beliefs to the interval endpoint in the direction that is most damaging to the agent's action choice. This structure delivers a sharp phase transition. If alignment is below one half, the optimal interval collapses to the prior and the agent ignores the adviser. If alignment is above one half, there is a unique nontrivial trust interval, and it expands monotonically with alignment, approaching full trust as alignment approaches certainty.

When the strategy space is binary (i.e., two actions and no additional private information), the robust solution becomes even more stark, in that the optimal trust region is bang-bang: either every recommendation is trusted or none is. The boundary between these regimes is a single cutoff alignment probability that depends only on two aggregate statistics of the adviser's information, namely the expected absolute gain from taking the second action when it is optimal and the expected absolute loss from taking it when it is suboptimal. The cutoff is minimized when gains and losses equal each other, and it increases as the decision becomes more asymmetric, capturing the idea that one-sided high-stakes mistakes require stricter standards of trust.

2 MODEL

Our formal model allows certain spaces to be finite or infinite. Whenever we work with an infinite space, we endow it with the Borel σ-algebra, and require all sets 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.

A state ω is drawn from a finite state space Ω, $|\Omega| = N$, according to a full-support prior distribution $\mu_0 \in \Delta(\Omega)$. An adviser observes partial information about ω, captured by a signal s whose distribution is pinned down by a signal function $\pi : \Omega \to \Delta(S)$. We will identify the adviser's information with the posterior about the state that a signal realization induces; let $S = \Delta(\Omega)$ and renormalize so that s is equal to the posterior about ω induced by s. Let τ denote the unconditional distribution of the adviser's posteriors s, with $M = \mathrm{supp}(\tau)$.

An agent takes an action $a \in A$, where A is a compact metric set. The agent observes a private type $\theta \in \Theta$, where Θ is a compact metric set, that captures the agent's own information about ω and her preferences, distributed according to a signal function $f : \Omega \to \Delta(\Theta)$. We assume that, conditional on the state, s and θ are distributed independently. The agent's ex-post payoff is given by a utility function $u(a; \omega, \theta)$, assumed bounded and continuous in a.

The adviser can send messages $m \in \Delta(\Omega)$ to the agent (without loss of generality, we set the message space to be the space of posteriors about the state). The agent can choose any strategy σ that prescribes how to act for any combination of the adviser's message and the agent's type, $\sigma : \Delta(\Omega) \times \Theta \to \Delta(A)$. Denote the set of all such strategies by Σ.

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

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 possible strategy σ according to its worst-case payoff

$$U(\sigma) \triangleq \alpha\, \mathbb{E}_{\mathrm{id},\sigma}[u(a; \omega, \theta)] + (1 - \alpha) \inf_{\beta \in B} \mathbb{E}_{\beta,\sigma}[u(a; \omega, \theta)], \tag{1}$$

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. 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).[2]

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

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

3 MAIN RESULTS

3.1 TRUST REGION STRATEGIES

In what follows, it will be convenient to separate dependence of the agent's strategy on the adviser's message and the agent's private information. To this end, we call a *private strategy* $\hat{\sigma}$ the mapping from types to actions $\hat{\sigma} : \Theta \to \Delta(A)$ that specifies how the agent uses her private information. If the agent has belief μ about the state and uses a private strategy $\hat{\sigma}$, then her expected payoff is:

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

[1] It can be shown that the assumption of truthful reporting of the belief is equivalent (in terms of equilibrium payoff consequences) to assuming that the aligned adviser is attempting to maximize the agent's expected payoff.

[2] Without loss of generality, adversarial strategies only use messages in M, since any message $m \notin M$ cannot be sent by an aligned adviser and hence reveals that the adviser is misaligned.

where the expectation is taken with respect to the conditional distribution of θ and the distribution of agent's actions induced by $\hat{\sigma}$. 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 that belief: $\hat{\sigma} \in \arg\max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$. The agent's strategy can be viewed as a specification of a private strategy for each possible message received from the adviser, $\sigma \sim (\hat{\sigma}(m))_{m \in \Delta(\Omega)}$.

Definition 1. $\sigma \sim (\hat{\sigma}(m))_{m \in \Delta(\Omega)}$ is a trust region strategy (TRS) if there exists a compact set $T \subset \Delta(\Omega)$ 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 $P(m) \in \text{argmax}_{m' \in T} U(\hat{\sigma}(m'), m)$.

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 posterior about the state. If a message m does not belong to the trust region T, the agent maps m to the trust region by acting *as if* her belief about the state were $P(m) \in T$; $P(m)$ is chosen to maximize—over all beliefs in the trust region—the agent's expected payoff under distribution m when the action is taken to be optimal for $P(m)$.

To provide further intuition, with slight abuse of notation, let $U(\mu) \triangleq \max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$ be the payoff to the agent when she takes a Bayes-optimal action at belief μ. Note that $U(\mu)$ is a convex function on $\Delta(\Omega)$; moreover, it is differentiable on the interior of the belief simplex if there exists a unique Bayes-optimal private strategy $\hat{\sigma}_0(\mu)$ at every belief μ. In that case, we can define $\nabla U(\mu)$—the gradient of the indirect payoff function treated as a function on $\mathbb{R}^N$—which maps each belief μ into the N-dimensional vector of state-contingent payoffs (associated with the Bayes-optimal strategy).[3] In particular, $U(\mu) = \nabla U(\mu) \cdot \mu$, where $\cdot$ denotes the standard dot product in $\mathbb{R}^N$. Moreover, for a TRS σ, we have $m' \in T$, $U(\hat{\sigma}(m'), m) = \nabla U(m') \cdot m$. Thus,

$$\text{argmax}_{m' \in T} U(\hat{\sigma}(m'), m) = \text{argmin}_{m' \in T} \underbrace{U(m) - U(m') - \nabla U(m') \cdot (m - m')}_{D_U(m, m')}.$$

The expression $D_U(m, m')$ is called the *Bregman distance* (associated with function U) between beliefs m and m'. Thus, under a TRS, messages outside of the trust region T are mapped into the closest belief—in the Bregman distance—in the trust region T. In particular, $P(m)$ always lies on the visible part (from the perspective of point m) of the boundary of T.[4]

3.2 Optimality of trust region strategies

We call two agent's strategies *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). *Any optimal strategy σ^* is equivalent to a trust region strategy with a connected trust region T.*

Proof. See Appendix A.1. □

Theorem 1 states that any optimal strategy can be interpreted as a TRS for some connected trust region T. Notice that if the adviser is always aligned, a TRS with the trust region equal to the entire belief space is trivially optimal. On the other extreme, if the adviser is always misaligned, the optimal TRS has a trust region equal to the prior—the agent always ignores the message of the adviser. In Section 3.4, we explore conditions under which the trust region can be guaranteed to be non-trivial.

In general, it is difficult to pin down the optimal trust region in closed form. A trade-off is created by two opposing forces: When the trust region expands, the expected payoff of the agent weakly increases conditional on the adviser being aligned but weakly decreases conditional on the adviser

[3] Formally, to define the gradient, we extend the function U beyond the probability simplex by assuming that U is homogeneous of degree 1 in the direction orthogonal to the simplex.

[4] Point $m' \in T$ is visible from m if the line segment connecting m' and m does not intersect $T \setminus \{m'\}$.

being misaligned. In Section 4.1, we study the binary-state case, in which the trust region is simply an interval, and some higher-dimensional examples that explore symmetry of the problem.

The trust region may fail to be convex. Intuitively, convexifying the trust region by adding a line segment connecting two beliefs may induce certain types of the misaligned adviser to report these additional beliefs (which may outweigh the benefits coming from aligned reports). However, the trust region can be taken to be a connected set. In fact, our proof establishes a slightly stronger property, effectively showing that the trust region is *convex in dual coordinates*, i.e., in the space of state-contingent payoffs. Intuitively, every belief has a corresponding state-contingent payoff induced by a Bayes-optimal action for that belief (as noted earlier, the state-contingent payoff is given by the gradient of the indirect utility function $U(\mu)$ at beliefs μ at which the optimal action is unique). The payoff of the misaligned adviser is non-linear in the reported belief but it is *linear* in the induced state-contingent payoff. In fact, the misaligned adviser with belief μ attempts to minimize $\mu \cdot w$ over all state-contingent payoffs w that some belief in the trust region induces. Thus, the set of state-contingent payoffs (induced by beliefs in the trust region) can be convexified. Convexification of the set of induced state-contingent payoffs connects the set of beliefs in the trust region.

Recalling our earlier definition of Bregman distance, note that we can equivalently think of the misaligned adviser with belief μ as choosing a message $m \in T$ to maximizes Bregman distance between μ and m; in particular, the message m must lie on the boundary of the trust region. A consequence is that adding non-boundary points to a trust region can only weakly increase the agent's payoff. Formally, we say that a set $A \subset \mathbb{R}^N$ is *non-hollow* if it contains all points $x \in \mathbb{R}^N$ with the property that every line going through x intersects A on both sides of x.

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

Note that being non-hollow is implied by convexity but not by connectedness. An example of a connected but hollow set is a sphere. If the trust region of some TRS is a sphere, then we can expand the trust region to the corresponding ball, since the misaligned adviser will never send messages in the interior of the ball.

As should be clear from our discussion, the trust region is typically not unique. Our results emphasized that the trust region can be taken to be a relatively large set. Note, however, that when the support M of the adviser's beliefs is finite, it is possible to construct an optimal discrete trust region T with $|T| \leq |M|$. Intuitively, at most one belief in the trust region is needed per every possible belief of the aligned adviser.[5] 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, $M = \Delta(\Omega)$ (see Section 4.1).

3.3 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, neither the commitment assumption nor the timing of moves 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 $\beta \in B$ of the adversarial adviser,[6] we let $\mathbb{P}_\beta(\cdot|m)$ denote the agent's conditional belief over the state Ω induced by message m given the adviser's strategy.

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

The condition of rationalizability means that the agent does not require commitment to follow the strategy—the agent can imagine the misaligned adviser pursuing an adversarial policy such that, after any message, the agent's strategy prescribes behaving myopically.

Theorem 2 (Robustly Rationalizable Solution). *Any robustly rationalizable strategy is optimal. If M and Θ are finite, a robustly rationalizable strategy exists.*

Proof. See Appendix A.2. □

[5] We emphasize, however, that it is not without loss of generality to assume that $T \subseteq M$.

[6] Without loss of generality, we assume that β uses only messages in M.

Assuming finite support of beliefs, Theorem 2 implies that there exists an optimal strategy β^* for the misaligned adviser such that the agent's optimal strategy is to simply best-respond to each posterior belief (computed via Bayes' rule given the strategy β^*) induced by on-path messages m.[7] In light of Theorem 1, such an equilibrium can still be taken to be a TRS. Treating the problem as a zero-sum game between the agent and the misaligned adviser, we will call $(\sigma^\star, \beta^\star)$ a trust region equilibrium (TRE) if $\sigma^\star$ is a TRS that is robustly rationalizable against the adversarial strategy $\beta^\star$.

In a TRE, messages $m \in M$ in the trust region are taken at face value because they are only reported by the aligned adviser (thus, Bayes' rule implies that $\mathbb{P}_{\beta^\star}(\cdot|m) = m$). 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 mapping to the boundary of the trust region defined in Theorem 1. Messages $m \notin M$ are sent with probability zero. In other words, the mapping from messages to beliefs induced by *(i)* Bayes' rule and *(ii)* minimizing Bregman distance to the trust region, coincide on the equilibrium path of a TRE. In Section 4.1, we use this structural property to characterize the trust region in a binary-state setting.

Theorem 2 implies that our problem is equivalent to a *constrained persuasion problem* for the misaligned adviser. When the misaligned adviser moves first, she is effectively choosing a Bayes-plausible distribution of posteriors for the agent subject to the constraint that the signal must be truthful in every state with probability at least α (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. This perspective will be useful in deriving bounds on the alignment probability α below which no information is communicated in any TRE (see Section 3.4).

From a policy perspective, Theorem 2 implies that implementing the optimal strategy does not require commitment by the agent. There exists a consistent conjecture about the form of misalignment that justifies the optimal strategy *ex-post*. That is, under that conjecture, the agent can simply observe the adviser's message, update her beliefs using Bayes' rule (using the conjecture about the misaligned adviser's strategy), and then take the optimal action for the resulting posterior.

Finally, from a technical perspective, Theorem 2 provides a practical way of certifying the solution optimality in applications (even with infinite belief and message spaces). To construct an optimal solution, it is sufficient 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. We use this method in Section 4.1.

3.4 Minimal Viable Alignment

In this section, we derive bounds on the alignment probability α above which informative communication can be supported between the adviser and the agent. In other words, we ask when the trust region used by the agent is non-trivial.

Formally, define the value of an adviser as

$$V \triangleq U^* - U_0,$$

where $U_0 \triangleq \sup_{\sigma \in \Sigma} \mathbb{E}_\sigma[u(a; \omega, \theta)]$ is the agent's optimal payoff in the absence of the adviser. Since the agent can always ignore adviser's messages, this value is non-negative, $V \geq 0$. We ask when this value is strictly positive, $V > 0$.

To answer this question, we assume that the adviser has a finite-support belief, $|M| = K < \infty$, and derive a bound on α that is independent of the agent's problem. Intuitively, if α is small enough, the misaligned adviser can use a stratregy β that effectively "jams" the signal created by truthful reporting of the aligned adviser. In such a case, the distribution of posteriors $\mathbb{P}_\beta(\cdot|m)$ held by the agent is degenerate: the trust region contains only the prior. 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 ($U(\mu)$ is *strictly* convex in the relevant range), V must be strictly positive.

[7] The assumption of finite M and Θ are made for technical reasons; verifying the assumptions of Sion (1958)'s minimax theorem (in particular, its continuity requirements) is difficult for a cheap-talk-like game with infinite-dimensional strategy spaces since the impact of messages on payoffs is entirely endogenous.

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

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

$$R(\tau) \triangleq \operatorname{rank} (\mu)_{\mu \in \operatorname{supp} \tau}. \tag{4}$$

Roughly, R captures the richness of the adviser's information: adviser's posteriors are located in an $(R-1)$-dimensional subspace of the $(N-1)$-dimensional belief simplex $\Delta(\Omega)$. For any τ, $R(\tau) \leq \min\{K, N\}$. The rank weakly decreases when the adviser's information is garbled. We assumed that the adviser has some information, $K \geq 2$, so $R(\tau) \geq 2$.

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

Proof. See Appendix A.3. □

The proof of Theorem 3 shows that, for any given τ, $\mathrm{MVA}(\tau)$ can be computed as the solution to a finite linear program. We establish bounds on this solution and then show that these bounds are tight by explicitly constructing adviser information structures that attain every MVA within the admissible range. In fact, the proof yields the stronger statement that, for any τ, $\mathrm{MVA}(\tau) \in \left[1/R(\tau), 1/2\right]$.

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 $\alpha < 1/2$, then the agent cannot benefit from the adviser: $N = 2$ and hence $\mathrm{MVA}(\tau) = 1/2$. In higher-dimensional problems, the adviser can be valuable at much lower alignment. In particular, if $R(\tau) = K = N$, then it suffices that $\alpha > 1/N$.

4 Trust Region Characterization

4.1 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). The belief is effectively one-dimensional: With slight abuse of notation, let $\mu \in [0, 1]$ denote the probability of state $\omega = 1$. For expositional clarity, we further assume that the agent's indirect payoff function $U(\mu) = \max_{\hat{\sigma}} U(\hat{\sigma}, \mu)$ is strictly convex and twice differentiable, and the adviser's posterior is distributed over $M = [0, 1]$ with a strictly positive probability density $\tau(\mu)$.[8] In this case, each $\mu \in [0, 1]$ can be associated with a unique Bayes-optimal private strategy $\hat{\sigma}_0(\mu)$.

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

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

If $\underline{\mu} = \bar{\mu}$, then the agent effectively ignores the adviser and optimally plays according to the prior, implying $\underline{\mu} = \bar{\mu} = \mu_0$. If $\underline{\mu} < \bar{\mu}$, then the agent plays according to $\hat{\sigma}_0(\underline{\mu})$ if $m \leq \underline{\mu}$ and according to $\hat{\sigma}_0(\bar{\mu})$ if $m \geq \bar{\mu}$.

Recall that the adversarial strategy of the misaligned adviser induces a posterior belief from the trust region that maximizes the Bregman distance from his true posterior; when the trust region is an interval—and hence its boundary consists of the two endpoints—the adversarial strategy admits a simple threshold characterization:

[8] The strictly convex indirect payoff function can be a result of the agent having a continuum of actions or, as we show in Appendix A.4, finitely many actions and a continuum of private types.

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

$$b(\underline{\mu}, \bar{\mu}) = \frac{\int_{\underline{\mu}}^{\bar{\mu}} \mu U''(\mu) d\mu}{\int_{\underline{\mu}}^{\bar{\mu}} U''(\mu) d\mu}. \tag{5}$$

Proof. See Appendix A.5. □

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 action at each posterior belief induced by the adviser's strategy. A necessary condition is that the average posterior belief induced by messages $m \leq \underline{\mu}$ is exactly $\underline{\mu}$, and the average posterior belief induced by messages $m \geq \bar{\mu}$ is exactly $\bar{\mu}$:[9]

$$\frac{\alpha \int_0^{\underline{\mu}} \mu\tau(\mu)d\mu + (1-\alpha) \int_{b(\underline{\mu},\bar{\mu})}^1 \mu\tau(\mu)d\mu}{\alpha \int_0^{\underline{\mu}} \tau(\mu)d\mu + (1-\alpha) \int_{b(\underline{\mu},\bar{\mu})}^1 \tau(\mu)d\mu} = \underline{\mu}, \tag{6}$$

$$\frac{\alpha \int_{\bar{\mu}}^1 \mu\tau(\mu)d\mu + (1-\alpha) \int_0^{b(\underline{\mu},\bar{\mu})} \mu\tau(\mu)d\mu}{\alpha \int_{\bar{\mu}}^1 \tau(\mu)d\mu + (1-\alpha) \int_0^{b(\underline{\mu},\bar{\mu})} \tau(\mu)d\mu} = \bar{\mu}. \tag{7}$$

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

Proposition 1 (Trust Region). *An optimal strategy exists; it is unique and robustly rationalizable. Its trust region, $T = [\underline{\mu}, \bar{\mu}]$, is equal to the prior $\{\mu_0\}$ when $\alpha \leq 1/2$; otherwise, it is defined by the unique solution to the system (6)-(7) that satisfies $\underline{\mu} \leq \mu_0 \leq \bar{\mu}$.*

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 is quite complex in a TRE. By Lemma 1, the misaligned adviser with belief $\mu \geq b(\underline{\mu}, \bar{\mu})$ is indifferent between sending all messages $m \leq \underline{\mu}$ since they all result in the same Bayes-optimal action $\hat{\sigma}_0(\underline{\mu})$. 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 $\beta^\star$ of the misaligned adviser must be such that *every* message $m \leq \underline{\mu}$ induces the posterior belief $\underline{\mu}$ via Bayes' rule. Since all messages $m \leq \underline{\mu}$ are sent on equilibrium path (the aligned adviser simply reports his belief truthfully), $\beta^\star$ must carefully break the misaligned adviser's indifference over these messages so that each of them induces the same posterior belief.

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 should result in more trust. The next result confirms this for the binary-state case:

Proposition 2 (Change in Alignment). *When $\alpha \geq 1/2$, $\underline{\mu}(\alpha)$ is strictly and continuously decreasing in α and $\bar{\mu}(\alpha)$ is strictly and continuously increasing in α. At $\alpha = 1/2$, $[\underline{\mu}, \bar{\mu}] = [\mu_0, \mu_0]$. At $\alpha = 1$, $[\underline{\mu}, \bar{\mu}] = [0, 1]$.*

Proof. See Appendix A.7. □

[9] 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.

4.2 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, $u(a_1; \omega) = 0$, 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 $\tau \in \Delta(\Delta(\Omega))$.

It is useful to define the absolute losses and gains from taking the second action relative to the first one:

$$L(\hat{\tau}) = \int_{-\infty}^{0} (-v)\, \hat{\tau}(dv), \quad G(\hat{\tau}) = \int_{0}^{+\infty} v\, \hat{\tau}(dv). \tag{8}$$

Also, define the following threshold:

$$\hat{\alpha}(\hat{\tau}) = \frac{\max\{L(\hat{\tau}), G(\hat{\tau})\}}{L(\hat{\tau}) + G(\hat{\tau})}. \tag{9}$$

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

Assumption 1 (Genericity). $L(\hat{\tau}) > 0, G(\hat{\tau}) > 0, L(\hat{\tau}) \neq G(\hat{\tau}), \tau(\{\mu : v(\mu) = 0\}) = 0.$

Proposition 3 (Binary Action Solution). *Suppose $A = \{a_1, a_2\}$, $|\Theta| = 1$, and Assumption 1 holds. If $\alpha \neq \hat{\alpha}(\hat{\tau})$, then the optimal solution exists, is unique, and is robustly rationalizable. In particular, 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 3, generically, the optimal solution is remarkably stark: either all adviser's messages are trusted or none do. As we saw previously, this bang-bang nature is not general and driven by the coarseness of the private strategies.

Notably, only the aggregate statistics $L(\hat{\tau})$ and $G(\hat{\tau})$ matter, not the detailed distribution of relative payoffs $\hat{\tau}$. The threshold $\hat{\alpha}(\hat{\tau})$ is minimized when $L(\hat{\tau}) = G(\hat{\tau})$, in which case $\mathrm{MVA} = 1/2$. Therefore, if $\alpha < 1/2$, the agent would never trust adviser irrespectively of $\hat{\tau}$. In contrast, for a given $\alpha > 1/2$, the trust condition $\hat{\alpha}(\hat{\tau}) < \alpha$ translates to L, G belonging to a cone defined by two linear conditions: $(1-\alpha)L \leq \alpha G$ and $(1-\alpha)G \leq \alpha L$. That is, the adviser is useful in a binary decision problem only if the expected gains and losses of one action relative to the other are not too distinct.

5 CONCLUSION

We studied robust decision-making when an agent relies on an informed adviser who may be misaligned. We characterized the agent's max–min optimal inference-and-action rule and showed that every optimal policy is equivalent to a trust-region strategy in belief space: the agent limits exposure to manipulation while preserving value from moderately informative advice. We also showed that optimal behavior is robustly rationalizable and identified when advice is robustly valuable. Finally, we fully characterized the optimal strategy in settings with a binary action or a binary state.

Our analysis suggests several directions for future work that lie beyond the scope of this paper. First, it would be useful to sharpen the comparative statics of the trust region with respect to the decision problem, the adviser's informativeness, and alignment. Beyond monotonicity in alignment, a central goal is to understand how the geometry of the trusted set evolves: which directions in belief space become trusted as alignment increases, and how greater informativeness can simultaneously raise the upside under truthful advice and expand adversarial leverage. Second, a natural direction is to characterize the trust region in richer environments, where it may take a more complex form; symmetric environments offer a promising route. Finally, it seems important to develop tractable computational methods and to study how the proposed solution can be integrated into modern AI systems as an interpretable layer between neural network outputs and downstream decision-making. All of these directions appear feasible by building on the present framework.

REFERENCES

David Blackwell. Comparison of experiments. In Jerzy Neyman (ed.), *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, pp. 93–102. University of California Press, Berkeley and Los Angeles, 1951.

Laura Doval and Alex Smolin. Persuasion and welfare. *Journal of Political Economy*, 132(7): 2451–2487, 2024.

David Gale and Hukukane Nikaido. The jacobian matrix and global univalence of mappings. *Mathematische Annalen*, 159(2):81–93, 1965.

Alex Gershkov, Benny Moldovanu, and Xianwen Shi. Order independence in sequential, issue-by-issue voting. *Mathematics of Operations Research*, 50(3):1635–1653, 2025.

Maurice Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

Przemysław Wojtaszczyk. *Banach Spaces for Analysts*, volume 25 of *Cambridge Studies in Advanced Mathematics*. Cambridge University Press, Cambridge, UK, 1991.

A PROOFS

A.1 PROOF OF THEOREM 1

We begin with a key lemma.

Lemma 2. *Any optimal solution σ^* is equivalent to an optimal solution that uses Bayes-optimal private strategies for all $m \in M$.*

Proof. Consider the set of state-contingent payoff profiles that are feasible for the agent (cf. Doval & Smolin (2024)):

$$W = \{\mathrm{w} \in \mathbb{R}^N : \exists\, \hat{\sigma},\ \mathrm{w}(\omega) = \mathbb{E}_{\hat{\sigma}}[u(a;\omega,\theta)|\omega],\ \forall\, \omega \in \Omega\}.$$

Since θ and s are conditionally independent, if the adviser has posterior s and the agent plays a private strategy that corresponds to payoff profile w, the resulting agent's expected payoff is $\mathrm{w} \cdot s$.

The set W is convex, because a convex combination of the private strategies delivers a convex combination of their respective payoff profiles. The set W is compact, because for any $\pi \in \mathbb{R}^{|\Omega|}$, $\max_{\mathrm{w}\in W} \pi^T \mathrm{w}$ exists and attained by some $\mathrm{w} \in W$ by the boundedness and continuity of u in a and the measurable maximum theorem.

Denote the (weak) Pareto frontier of W by W^P:

$$W^P = \{\mathrm{w} \in W : \nexists\, \mathrm{w}' \in W, \forall\, \omega \in \Omega, \mathrm{w}'(\omega) > \mathrm{w}(\omega)\}.$$

Since W is convex and compact, by the supporting hyperplane theorem, a private strategy $\hat{\sigma}$ is Bayes-optimal (for some belief) if and only if it delivers a payoff profile in W^P. Therefore, if $\hat{\sigma}$ is not Bayes-optimal, there exists a dominating $\hat{\sigma}'$ (which can be taken to be Bayes-optimal itself) such that for all $\omega \in \Omega$, $\mathbb{E}_{\hat{\sigma}'}[u(a;\omega,\theta)|\omega] > \mathbb{E}_{\hat{\sigma}}[u(a;\omega,\theta)|\omega]$.

Take σ^* and, for every message $m \in M$, if $\hat{\sigma}^*(m)$ is not Bayes-optimal for some belief, replace it with a Bayes-optimal dominating strategy $\hat{\sigma}'(m)$. The new strategy must still be optimal. Indeed, the agent's payoff is

$$\mathbb{E}_{\mu\sim\tau}[\alpha\, \mathrm{w}(\hat{\sigma}(\mu)) \cdot \mu + (1-\alpha) \inf_{m\in M} \{\mathrm{w}(\hat{\sigma}(m)) \cdot \mu\}],$$

which pointwise increases after the dominating change.

Hence, the new strategy σ_0 is also optimal. Moreover, the expected payoff must stay the same since σ^* was optimal to begin with; in particular, σ_0 makes changes to the agent's strategy only for messages m that have joint probability zero in equilibrium. Thus, σ^* is equivalent to σ_0. □

We can now finish the proof of Theorem 1. Pick any optimal solution σ^*. By Lemma 2, σ^* is equivalent to an optimal strategy that uses only Bayes-optimal private strategies. Denote by Σ_0 the set of those private strategies, and let T_0 be the closure of set of beliefs at which those private strategies are Bayes-optimal. (By continuity, the closure doesn't affect the strategy's payoff.)

Observe that the agent's payoff coming from the misaligned adviser depends only on the set Σ_0. Thus, the mapping from messages to the private strategies in Σ_0 must maximize the expected payoff conditional on the adviser being aligned. Since the aligned adviser is non-strategic, maximization can be performed pointwise, message by message (without loss of optimality, also for messages that are sent with probability zero by the aligned adviser). In particular, for $m \in T_0$, we can set $\hat{\sigma}^*(m)$ to be the Bayes-optimal strategy for m; for $m \notin T_0$, we can set $\hat{\sigma}^*(m) = \hat{\sigma}^*(P(m))$ where $P(m) \in \arg\max_{m' \in T_0} U(\hat{\sigma}^*(m'), m)$. This way we have constructed a TRS (with the trust region T_0) that is equivalent to σ^*—and is hence optimal.[10]

We now show that for any optimal TRS σ^*, the trust region T_0 can be enlarged (while preserving the payoffs) to a connected trust region T_1. Assume that T_0 is not connected and take any $m_1, m_2 \in T_0$ that belong to different connected components of T_0: $m_1 \in T_0^1$ and $m_2 \in T_0^2$. Consider the welfare profiles $\mathrm{w}_1 \triangleq \mathrm{w}(\hat{\sigma}^*(m_1))$ and $\mathrm{w}_2 \triangleq \mathrm{w}(\hat{\sigma}^*(m_2))$ induced by those messages in the considered solution. Define the subset of Pareto optimal welfare profiles that dominate some weighted average of those profiles $\mathrm{w}(\sigma) \triangleq \gamma \mathrm{w}_1 + (1-\gamma)\mathrm{w}_2$:

$$W^D(m_1, m_2) = \{\mathrm{w} \in W^P : \exists\, \gamma \in [0,1], \mathrm{w} \geq \mathrm{w}(\gamma)\}.$$

Consider any $\mathrm{w} \in W^D(m_1, m_2)$ that dominates $\mathrm{w}(\gamma)$ for some γ. Since $\mathrm{w} \in W^P$, w is generated by a private strategy $\hat{\sigma}(\mathrm{w})$ Bayes-optimal at a set of beliefs $M_1(\mathrm{w})$. We enlarge T_0 by including to it $M_1(\mathrm{w}) \setminus T_0$ together with the prescription to play $\hat{\sigma}(\mathrm{w})$ at those messages. Doing so doesn't decrease the payoff from the misaligned adviser, because he could already send messages m_1 and m_2. However, it weakly increases the payoff from the aligned adviser by giving her more opportunities to choose from.

Since W is convex, $W^D(m_1, m_2)$ is connected. Furthermore, $M_1(\mathrm{w})$ is upper-hemicontinuous with connected (convex) values by being a normal-cone correspondence. Therefore, the union $\bigcup_{\mathrm{w} \in W^D(m_1,m_2)} M_1(\mathrm{w})$ is connected and contains m_1 and m_2. Thus, doing this change for all $\mathrm{w} \in W^D(m_1, m_2)$, connects the components T_0^1 and T_0^2, with the trust region weakly expanding. Doing this change for all connected components of T_0 results in a connected set T_1.

Finally, by continuity of payoffs, we can without loss of generality consider the closure of the set of used private strategies, and hence the trust region can be chosen to be equal to $T = \mathrm{cl}\, T_1$, which is a compact and connected subset of $\Delta(\Omega)$.

By construction, the new strategy σ_1 is optimal. Moreover, the expected payoff must stay the same since σ^* was optimal to begin with; therefore, the new strategy makes changes to the agent's strategy only for messages that have joint probability zero in equilibrium. Thus, σ^* is equivalent to σ_1.

A.2 Proof of Theorem 2

Suppose M and Θ are finite. For any given strategies of the misaligned adviser and the agent, (β, σ), the agent's payoff is, with a slight overload of notation for U,

$$U(\beta, \sigma) \triangleq \alpha \sum_{\mu \in M, \omega \in \Omega, \theta \in \Theta} \tau(\mu)\mu(\omega) f(\theta|\omega) \int_A u(a; \omega, \theta)\sigma(da|\mu, \theta) + \\ (1-\alpha) \sum_{\mu, m \in M, \omega \in \Omega, \theta \in \Theta} \tau(\mu)\mu(\omega)\beta(m|\mu) f(\theta|\omega) \int_A u(a; \omega, \theta)\sigma(da|m, \theta).$$

Clearly, B and Σ are convex. Since M is finite, $B = \times_{m \in M} \Delta(M)$ is compact. Since, M and Θ are finite and $\Delta(A)$ is equipped with the weak topology, $\Sigma = \times_{m \in M, \theta \in \Theta} \Delta(A)$ is compact. $U(\beta, \sigma)$ is affine in β and in σ; therefore it is concave-convexlike in Sion (1958)'s terminology. For each $\sigma \in \Sigma$, $U(\beta, \sigma)$ is continuous in β. For each $\beta \in B$, $U(\beta, \sigma)$ is continuous in σ. Therefore, a minimax

[10] It is equivalent to σ^* because it is weakly better than σ^* but σ^* was optimal.

theorem applies in its infsup variation (e.g., Theorem 4.2', Sion (1958)) and

$$\sup_{\sigma\in\Sigma}\inf_{\beta\in B} U(\beta,\sigma) = \inf_{\beta\in B}\sup_{\sigma\in\Sigma} U(\beta,\sigma).$$

Furthermore, for any given β, $\phi(\beta) \triangleq \sup_{\sigma\in\Sigma} U(\beta,\sigma)$ is attained because Σ is compact and $U(\beta,\sigma)$ is continuous in σ. Similarly,, for any given σ, $\psi(\sigma) \triangleq \inf_{\beta\in B} U(\beta,\sigma)$ is attained because B is compact and $U(\beta,\sigma)$ is continuous in β. Because, $U(\beta,\sigma)$ is continuous, $\phi(\beta)$ is lower-semicontinuous and $\psi(\sigma)$ is upper-semicontinuous. Thus, we can set $\sigma^* \in \arg\max_{\sigma\in\Sigma}\psi(\sigma)$ and $\beta^* \in \arg\min_{\beta\in B}\phi(\beta)$ and they form a saddle point:

$$U(\beta^*,\sigma) \le U(\beta^*,\sigma^*) \le U(\beta,\sigma^*), \quad \forall\,\beta\in B, \sigma\in\Sigma. \tag{10}$$

Therefore, β^* is an adversarial adviser's strategy to σ^*, whereas σ^* is a best-response of the agent to β^*. The latter implies—since $\alpha > 0$ and all $m \in M$ are on-path—that after any $m \in M$, the private strategy $\hat{\sigma}^*(m)$ is Bayes-optimal given β^*, and hence σ^* is robustly rationalizable.

Conversely, for any M and Θ, any σ^* and β^* that satisfy the conditions of the theorem form a saddle point with property (10). Then, for any $\sigma \in \Sigma$:

$$U(\sigma) = \inf_{\beta\in B} U(\beta,\sigma) \le U(\beta^*,\sigma) \le U(\beta^*,\sigma^*) = \min_{\beta\in B} U(\beta,\sigma^*) = U(\sigma^*),$$

where the third comparison uses the saddle property and the fourth comparison uses the fact that β^* is adversarial to σ^*. Therefore, σ^* is an optimal solution.

A.3 Proof of Theorem 3

Notation: In this section, 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, and by $X^\top$ a transpose of a matrix X.

Since μ_0 has full support, we can equivalently identify 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, $\operatorname{rank}\Pi = R$.[11]

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 garbling of his information:

$$G(B) \triangleq \alpha I_K + (1-\alpha)B. \tag{11}$$

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 $\mathrm{MVA}(\Pi)$.

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

$$D(\Pi)G(B) = D(\Pi)(\alpha I_K + (1-\alpha)B) = 0_{(N-1)\times K}, \tag{12}$$

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:

$$\Lambda(\Pi) = \max_{G\in\mathbb{R}^{K\times K},\,\alpha\in\mathbb{R}} \quad \alpha \tag{13}$$

$$\text{s.t.} \quad G \ge \alpha I_K,\ G1_K = 1_K, \tag{14}$$

$$D(\Pi)G = 0_{(N-1)\times K}. \tag{15}$$

[11] 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.

Lemma 3. $\mathrm{MVA}(\Pi) = \Lambda(\Pi)$.

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

$\Rightarrow$ For any given α, if B is such that $\Pi G(B)$ is Blackwell uninformative, then we showed that $G(B)$ must satisfy conditions (14-15). By the maximization nature of the problem, if $\alpha > \Lambda$, those conditions cannot be satisfied.

$\Leftarrow$ If $\alpha \leq \Lambda$, then there exists G that satisfies conditions (14-15) (e.g., the argmax). If $\Lambda = 1$, then B can be arbitrary. Otherwise, set $B = (G - \alpha I_K)/(1-\alpha)$. It is straightforward that the so-defined B is a stochastic matrix and by construction $\Pi G(B)$ is uninformative. □

Lemma 3 provides a computationally tractable characterization of MVA for any given Π and sets the stage for the rest of the proof, which we split into two lemmas.

Lemma 4. $\mathrm{MVA}(\Pi) \in [1/R(\Pi), 1/2]$. *If* $R(\Pi) = K$, *then* $\mathrm{MVA}(\Pi) = 1/K$.

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

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

If α and G satisfy (14-15), then $B = (G - \alpha I_K)/(1-\alpha)$ is a stochastic matrix and

$$D(\Pi)B = -\frac{\alpha}{1-\alpha} D(\Pi). \tag{16}$$

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

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

If $R = K$, then $K \leq N$ and $\mathrm{rank} D(\Pi) = K - 1$. Thus, $\mathrm{rank}\,\ker D(\Pi) = K - (K-1) = 1$ and, because $D1_K = 1_{N-1} - 1_{N-1} = 0_{N-1}$, $\ker D = \mathrm{span}\{1_K\}$. Thus, for (G, α) to satisfy (15), 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 (14), it must be that $\alpha \leq 1/K$. Thus, $\mathrm{MVA}(\Pi) \leq 1/K$.

At the same time, if $\alpha \leq 1/K$, then (G, α) satisfy (14-15) 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 to be uninformative about his estimate, not only about the state.) It follows, that $\mathrm{MVA} \geq 1/K$ and, therefore, $\mathrm{MVA}(\Pi) = 1/K$.

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

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[12]

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

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

$$W^\top X = I_{R-1}, \tag{17}$$

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

$$W^\top 1_K = 0_{R-1}. \tag{18}$$

[12] By the Auerbach theorem, there exist $v_1, \ldots, v_{R-1} \in \mathbb{W}$ and $\phi_1, \ldots, \phi_{R-1} \in \mathbb{W}^*$ such that $\|v_i\|_1 = 1$, $\|\phi_i\|_{\mathbb{W}^*} = 1$, and $\phi_i(v_j) = \delta_{ij}$ (Section II.E, Lemma 11 in Wojtaszczyk (1991); see also Gershkov et al. (2025) for another recent application). By 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, for $w \in \mathbb{W}$, $w_i^\top x_j = \tilde{\phi}_j(w_i) = \phi_j(w_i) = \delta_{ij}$.

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

$$\sum_{k=1}^{K} \bar{w}_k = \sum_{i=1}^{R-1} \|w_i\|_1 = R - 1. \tag{19}$$

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

$$B = \frac{1}{R-1}(1_K \bar{w}^\top - XW^\top). \tag{20}$$

Nonnegativity. For all j, k,

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

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

Stochasticity. By (18) and (19):

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

Uninformativeness. By (17) and (18):

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

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 (15) (see (16)). The result follows. □

Lemma 5. *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 (13).

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

$$(e_K^i - e_K^1)^\top G = 0, \quad i = 2, \ldots, K-2, \qquad (e_K^{K-1} - \delta e_K^1 - (1-\delta)e_K^K)^\top G = 0. \tag{21}$$

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 boil down 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

$$\alpha \le \alpha^\dagger \triangleq \frac{1}{2+\delta(K-3)}. \tag{22}$$

and thus $\mathrm{MVA}(\Pi) \le \alpha^\dagger$. Whenever $\delta \ge (K-4)/(K-3)$, $\alpha^\dagger \le 1/(K-2)$ 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 (21)):

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

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.

□

A.4 On Strictly Convex Indirect Utility

In this section, we show that the indirect utility is strictly convex in the case of full-support agent's private information.

Specifically, we assume that the agent's ex-post payoff is type-independent, $u(a;\omega)$ and identify θ with the belief it induces in the absence of any other information: $\Theta \subseteq \Delta(\Omega)$, $\theta(\omega) = \Pr(\tilde{\omega} = \omega|\theta)$. We denote by ν the final posterior that the agent forms, i.e., conditional on both the adviser's message and the agent's type:

$$\nu_{\mu,\theta} \triangleq \Pr(\omega|\mu,\theta) = \frac{\mu(\omega)f(\theta|\omega)}{\sum_{\omega'\in\Omega}\mu(\omega')f(\theta|\omega')}. \tag{23}$$

A necessary and sufficient condition for a private strategy $\hat{\sigma}$ to be Bayes-optimal at any given posterior μ, $\hat{\sigma} \in \arg\max_{\hat{\sigma}'} U(\hat{\sigma}', \mu)$, is that $\hat{\sigma}(\cdot|\theta) \in \Delta(A)$ is an optimal best-response with respect to $\nu_{\mu,\theta}$: for all $a \in \operatorname{supp}\hat{\sigma}(\cdot|\theta)$,

$$a \in \arg\max_{a'\in A} \sum_{\omega\in\Omega} \nu_{\mu,\theta}(\omega)u(a';\omega). \tag{24}$$

Assumption 2. A is finite and there exist $a_1, a_2 \in A$ and $\mu \in \mathrm{int}(\Delta(\Omega))$ such that $\mathbb{E}_\mu[u(a_1;\omega)] = \mathbb{E}_\mu[u(a_2;\omega)] > \mathbb{E}_\mu[u(a;\omega)]$ for all $a \notin \{a_1, a_2\}$. In addition, for each $\omega \in \Omega$ either $u(a_1;\omega) > u(a_2;\omega)$ or $u(a_2;\omega) > u(a_1;\omega)$.

Lemma 6. *Suppose θ has full support on $\Delta(\Omega)$ and Assumption 2 holds. Then, $U(\mu)$ is strictly convex in the interior of $\Delta(\Omega)$.*

Proof. A sufficient condition for strict convexity of $U(\mu)$ in the interior of $\Delta(\Omega)$ is that for any $\mu_1, \mu_2 \in \mathrm{int}(\Delta(\Omega))$, $\mu_1 \ne \mu_2$,

$$\arg\max_{\hat{\sigma}} U(\hat{\sigma}, \mu_1) \cap \arg\max_{\hat{\sigma}} U(\hat{\sigma}, \mu_2) = \emptyset.$$

Fix any such μ_1, μ_2. Let $\mu \in \mathrm{int}(\Delta(\Omega))$ be the belief from Assumption 2 and define $d(\omega) \triangleq u(a_1;\omega) - u(a_2;\omega)$, $r(\omega) \triangleq \mu_2(\omega)/\mu_1(\omega)$. By Assumption 2 and continuity of the expected payoff in belief, there exists an open neighborhood $O \subset \mathrm{int}(\Delta(\Omega))$ of μ such that for every $\nu \in O$, action a_1 is uniquely optimal whenever $\nu \cdot d > 0$, and not optimal whenever $\nu \cdot d < 0$, because it is outperformed by a_2. Define

$$R_1 \triangleq \{\nu \in O : \nu \cdot d > 0\}, \qquad R_2 \triangleq \{\nu \in \Delta(\Omega) : \nu \cdot d < 0\}.$$

Bayes' rule implies that for every ω and θ, $\nu_{\mu_2,\theta} = T(\nu_{\mu_1,\theta})$, where $T : \mathrm{int}(\Delta(\Omega)) \to \mathrm{int}(\Delta(\Omega))$ is the map defined by

$$T(\nu)(\omega) \triangleq \frac{\nu(\omega)r(\omega)}{\sum_{\omega'}\nu(\omega')r(\omega')}.$$

Since $\mu_1 \neq \mu_2$, r is not constant; because $d(\omega) \neq 0$ for all ω, the hyperplanes $\{\nu : \nu \cdot d = 0\}$ and $\{\nu : \nu \cdot (r * d) = 0\}$ (where "$*$" is the component-wise product) are distinct. As $\mu \in O \cap \{\nu : \nu \cdot d = 0\}$, we can choose $\bar{\nu} \in O$ such that $\bar{\nu} \cdot d = 0$ and $\bar{\nu} \cdot (r * d) \neq 0$. Without loss of generality, suppose $\bar{\nu} \cdot (r * d) < 0$ (otherwise swap the labels of a_1 and a_2). By continuity, there exists a nonempty open set $N \subset R_1$ such that $\nu \cdot (r * d) < 0$ for all $\nu \in N$. For every $\nu \in N$,

$$T(\nu) \cdot d = \frac{\nu \cdot (r * d)}{\nu \cdot r} < 0,$$

so $T(N) \subset R_2$. The map $\theta \mapsto \nu_{\mu_1,\theta}$ is continuous and onto $\mathrm{int}(\Delta(\Omega))$. Hence $\Theta_0 \triangleq \{\theta : \nu_{\mu_1,\theta} \in N\}$ is nonempty and open; since θ has full support on $\Delta(\Omega)$, it has strictly positive probability.

For every $\theta \in \Theta_0$ we have $\nu_{\mu_1,\theta} \in R_1$ and $\nu_{\mu_2,\theta} \in R_2$. This means that the private strategies optimal at μ_1 and μ_2 must necessarily differ on $\theta \in \Theta_0$. The result follows. □

A.5 Proof of Lemma 1

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,

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

and we assumed that U is strictly convex and twice differentiable. 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 $\mu' \in [\underline{\mu}, \bar{\mu}]$, $U(\hat{\sigma}(\mu'), \mu) \geq \min\{U(\hat{\sigma}(\underline{\mu}), \mu), U(\hat{\sigma}(\bar{\mu}), \mu)\}$. Thus, the misaligned adviser's strategy takes a threshold form. The threshold $b(\underline{\mu}, \bar{\mu})$ is the intersection point of the supporting lines to U at points $\underline{\mu}$ and $\bar{\mu}$:

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

Rearranging, we obtain:

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

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

A.6 Proof of Proposition 1

By Lemma 1 and Corollary 2, the choice of an optimal strategy for the agent boils down to optimization over the extreme points $\underline{\mu}$, $\bar{\mu}$ of the trust interval with the corresponding payoff:

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

The function $U(\underline{\mu}, \bar{\mu})$ is continuously differentiable with the partial derivatives at points with $\underline{\mu} < \bar{\mu}$:

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

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 posterior regions in which the aligned and misaligned advisers induce that action, weighted by the alignment parameter. (Terms involving $\partial b/\partial \underline{\mu}$ and $\partial b/\partial \bar{\mu}$ vanish because at $\mu = b(\underline{\mu}, \bar{\mu})$ the misaligned adviser is indifferent between the two messages.)

Whenever the trust region is non-singleton, $\underline{\mu} < \bar{\mu}$, at the optimal choice of $\underline{\mu}$ and $\bar{\mu}$ these partial derivatives must equal zero, $\partial U/\partial \underline{\mu} = 0$ and $\partial U/\partial \bar{\mu} = 0$. Since $U''(\cdot) > 0$, these first-order conditions can be rearranged as follows. Define functions Ψ_1 and Ψ_2 as

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

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

Then, $\Psi_1(\underline{\mu}, \bar{\mu}, \alpha) = \Psi_2(\underline{\mu}, \bar{\mu}, \alpha) = 0$ is equivalent to conditions (6) and (7).

First, we show that conditions (6) and (7) are incompatible with $\alpha < 1/2$. (If M was finite, this would follow directly from Theorem 3.) Indeed, if those conditions hold then (for the rest of the proof, we will often omit the arguments of the function b for brevity):

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

where the inequalities hold because $\underline{\mu} \leq b(\underline{\mu}, \bar{\mu}) \leq \bar{\mu}$ and the equality is a consequence of (6) and (7) (their sum). 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 $\alpha \geq 1/2$ the solution to (6) and (7) such that $\underline{\mu} \leq \bar{\mu}$ exists. (Note that at $\alpha = 1/2$, $[\underline{\mu}, \bar{\mu}] = [\mu_0, \mu_0]$ is a solution.) For the rest of this proof, we omit the dependence of Ψ_i on α. By Lemma 1 and direct inspection, $b(\underline{\mu}, \bar{\mu})$ is strictly and continuously increasing in its arguments, so $\Psi_1(\underline{\mu}, \bar{\mu})$ is strictly and continuously decreasing in $\underline{\mu}$ for each $\bar{\mu}$. Furthermore,

$$\Psi_1(0, \bar{\mu}) = (1-\alpha)\int_{b(0,\bar{\mu})}^1 \mu\tau(\mu)d\mu \geq 0,$$

$$\Psi_1(1, \bar{\mu}) = \alpha\int_0^1 (\mu - 1)\tau(\mu)d\mu < 0.$$

Therefore, for each $\bar{\mu}$, a best-response $b_1(\bar{\mu})$ such that $\Psi_1(b_1(\bar{\mu}), \bar{\mu}) = 0$ exists and is unique. Since $\Psi_1(\underline{\mu}, \bar{\mu})$ strictly decreases in $\bar{\mu}$, $b_1(\bar{\mu})$ strictly decreases in $\bar{\mu}$. Finally, for any $\bar{\mu}$,

$$\int_{b_1(\bar{\mu})}^1 (\mu - b_1(\bar{\mu}))\tau(\mu)d\mu \geq \int_{b(b_1(\bar{\mu}),\bar{\mu})}^1 (\mu - b_1(\bar{\mu}))\tau(\mu)d\mu \geq \int_0^{b_1(\bar{\mu})} (b_1(\bar{\mu}) - \mu)\tau(\mu)d\mu,$$

where the second inequality holds because $\alpha \geq 1/2$ and $\Psi_1(b_1(\bar{\mu}), \bar{\mu}) = 0$. Thus, for any $\bar{\mu}$, $b_1(\bar{\mu}) \leq \mu_0$.

Analogously, for each $\underline{\mu}$, a best-response $b_2(\underline{\mu})$ such that $\Psi_2(\underline{\mu}, b_2) = 0$, exists, unique, strictly decreases in $\underline{\mu}$, and is everywhere greater than μ_0.

Therefore, a solution to (6) and (7) is any $\underline{\mu} \in [0, \mu_0]$ and $\bar{\mu} = b_2(\underline{\mu}) \in [\mu_0, 1]$ 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 $D = \{(\underline{\mu}, \bar{\mu}) : \underline{\mu} \in [0, \mu_0], \bar{\mu} \in [\mu_0, 1]\}$. To establish uniqueness, consider the function $-\Psi = (-\Psi_1, -\Psi_2)$ on D. Any solution must satisfy $-\Psi(\underline{\mu}, \bar{\mu}) = (0, 0)$. Observe that for any $(\underline{\mu}, \bar{\mu}) \in D$,

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

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

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

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

Finally, we show that the proposed trust-region strategy is robustly rationalizable by explicitly constructing a TRE. For $\alpha \geq 1/2$, we need to construct a measurable strategy of the misaligned adviser $\beta : [0, b] \to [\bar{\mu}, 1]$ such that for every set $X \subseteq [\bar{\mu}, 1]$ with $\alpha\tau(X) + (1-\alpha)\tau(\beta^{-1}(X)) > 0$,

$$\frac{\alpha \int_X \mu\tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} \mu\tau(\mu)d\mu}{\alpha \int_X \tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} \tau(\mu)d\mu} = \bar{\mu}. \tag{25}$$

(The construction of $\beta : (b, 1] \to [0, \underline{\mu}]$ is analogous.) To this end, define two finite atomless nonnegative measures:

$$\begin{aligned}
\nu(Y) &\triangleq (1-\alpha)\int_Y (\bar{\mu}-\mu)\tau(\mu)d\mu, \quad Y \subseteq [0, b]\\
\eta(X) &\triangleq \alpha \int_X (\mu-\bar{\mu})\tau(\mu)d\mu, \quad X \subseteq [\bar{\mu}, 1].
\end{aligned}$$

Observe that condition (7) is precisely $\eta([\bar{\mu}, 1]) = \nu([0, b])$ whereas condition (25) is the pushforward identity:

$$\eta(X) = \nu(\beta^{-1}(X)), \quad X \subseteq [\bar{\mu}, 1].$$

In other words, we need to find β that transports ν to η. It is always possible. For a canonical quantile construction, define the cumulative mass functions $F_\nu(\mu) \triangleq \nu([0, \mu])$ for $\mu \in [0, b]$ and $F_\eta(\mu) \triangleq \eta([\bar{\mu}, \mu])$ for $\mu \in [\bar{\mu}, 1]$. The transport map can then be set:

$$\beta(\mu) = F_\eta^{-1}(F_\nu(\mu)), \quad \mu \in [0, b],$$

where $F_\eta^{-1}(\cdot)$ is the generalized inverse: $F_\eta^{-1}(q) = \inf\{\mu \in [\bar{\mu}, 1] : F_\eta(\mu) \geq q\}$.

For $\alpha < 1/2$, $T = \{\mu_0\}$, so the misaligned adviser is indifferent between all messages and it suffices to construct a strategy $\beta : [0, 1] \to [0, 1]$ such that for all $X \subseteq [0, 1]$ with $\alpha \int_X \tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} \tau(\mu)d\mu > 0$,

$$\frac{\alpha \int_X \mu\tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} \mu\tau(\mu)d\mu}{\alpha \int_X \tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} \tau(\mu)d\mu} = \mu_0, \tag{26}$$

which is equivalent to:

$$\alpha \int_X (\mu-\mu_0)\tau(\mu)d\mu + (1-\alpha)\int_{\beta^{-1}(X)} (\mu-\mu_0)\tau(\mu)d\mu = 0.$$

To do that, observe that $\int_0^1(\mu-\mu_0)\tau(\mu)d\mu = 0$ and $\int_0^{\mu_0}(\mu_0-\mu)\tau(\mu)d\mu = \int_{\mu_0}^1(\mu-\mu_0)\tau(\mu)d\mu > 0$. Since $\alpha \in (0,1/2)$, $\alpha/(1-\alpha) \in (0,1)$ and by the intermediary value theorem, there exist $\mu_L \in (0,\mu_0)$ and $\mu_H \in (\mu_0,1)$ such that

$$\int_0^{\mu_L}(\mu_0-\mu)\tau(\mu)d\mu = \frac{\alpha}{1-\alpha}\int_0^{\mu_0}(\mu_0-\mu)\tau(\mu)d\mu,$$
$$\int_{\mu_H}^1(\mu-\mu_0)\tau(\mu)d\mu = \frac{\alpha}{1-\alpha}\int_{\mu_0}^1(\mu-\mu_0)\tau(\mu)d\mu.$$

By construction,

$$\int_0^{\mu_L}(\mu_0-\mu)\tau(\mu)d\mu = \frac{\alpha}{1-\alpha}\int_{\mu_0}^1(\mu-\mu_0)\tau(\mu)d\mu, \tag{27}$$
$$\int_{\mu_H}^1(\mu-\mu_0)\tau(\mu)d\mu = \frac{\alpha}{1-\alpha}\int_0^{\mu_0}(\mu_0-\mu)\tau(\mu)d\mu, \tag{28}$$
$$\int_{\mu_L}^{\mu_H}(\mu-\mu_0)\tau(\mu)d\mu = 0. \tag{29}$$

We can set $\beta(\mu) = \beta_L(\mu)$ when $\mu \in [0,\mu_L]$, $\beta(\mu) = \mu_0$, when $\mu \in (\mu_L,\mu_H)$, and $\beta(\mu) = \beta_H(\mu)$, when $\mu \in [\mu_H,1]$. Here, β_L is a quantile transport map that transports measure $\nu_L(Y) = (1-\alpha)\int_Y(\mu_0-\mu)\tau(\mu)d\mu$ on $[0,\mu_L]$ to measure $\eta_L(X) = \alpha\int_X(\mu-\mu_0)\tau(\mu)d\mu$ on $[\mu_0,1]$, just like in the case of $\alpha \geq 1/2$; it ensures that (26) holds for all $X \subseteq (\mu_0,1]$. Similarly, β_H is a quantile transport map that transports measure $\nu_H(Y) = (1-\alpha)\int_Y(\mu-\mu_0)\tau(\mu)d\mu$ on $[\mu_H,1]$ to measure $\eta_H(X) = \alpha\int_X(\mu_0-\mu)\tau(\mu)d\mu$ on $[0,\mu_0]$; it ensures that (26) holds for all $X \subseteq [0,\mu_0)$. (The transported masses match the targets by (27) and (28).) Finally, by (29), (26) holds for $\mu = \mu_0$. The result follows.

A.7 Proof of Proposition 2

At $\alpha = 1/2$, $[\underline{\mu},\bar{\mu}] = [\mu_0,\mu_0]$ satisfies conditions (6) and (7). At $\alpha = 1$, $[\underline{\mu},\bar{\mu}] = [0,1]$ satisfies conditions (6) and (7).

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

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

As we argued in the proof of Proposition 1, for all $\underline{\mu}$, $\bar{\mu}$, $\alpha > 1/2$, $\det J(\underline{\mu},\bar{\mu},\alpha) > 0$, and therefore, by the implicit function theorem, optimal $\underline{\mu}(\alpha)$ and $\bar{\mu}(\alpha)$ are continuously differentiable and[13]

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

Consequently,

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

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

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

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

The result follows.

A.8 Proof of Proposition 3

With a small abuse of notation, we can parameterize each private strategy by $\hat{\sigma} = \Pr(a = a_2)$. Then, by the arguments behind Theorem 1, if the agent employs the set of private strategies $\hat{\Sigma}_0 = \{\hat{\sigma}(m)\}_{m \in M}$, 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:

$$\begin{aligned} &\int_{-\infty}^{0} (\alpha\hat{\sigma}_L + (1-\alpha)\hat{\sigma}_H) v \, \hat{\tau}(dv) + \int_{0}^{+\infty} (\alpha\hat{\sigma}_H + (1-\alpha)\hat{\sigma}_L) v \, \hat{\tau}(dv) \\ &= \hat{\sigma}_L((1-\alpha)G - \alpha L) + \hat{\sigma}_H(\alpha G - (1-\alpha)L). \end{aligned} \tag{30}$$

The optimal choice of $\hat{\sigma}_L$ and $\hat{\sigma}_H$ must maximize (30) 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:

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

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 nil 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 agent's strategy.

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