Friend or Foe: Delegating to an AI whose Alignment is Unknown *

Drew Fudenberg† Annie Liang‡

April 20, 2026

Abstract

We study delegation of a risky decision to an AI whose objective may be aligned with the designer's or misaligned against it. The designer can limit delegation by setting group-specific treatment rates and by choosing inputs with known bounds on conditional treatment success. We characterize the risk-reward frontier of achievable best- and worst-case payoffs. When the conditional treatment sucess is unrestricted, the frontier is a single line segment with slope -1, so any gain in best-case performance requires an equal loss in the worst case. When the designer can also choose limits on probability of treatment success given the inputs, the optimal policy has a simple asymmetric structure. In groups where treatment is ex ante unlikely to help, the designer never limits how conclusive a covariate can be about treatment success, but may limit how conclusive it can be about treatment harm; in groups where treatment is ex ante likely to help, the reverse holds. As the designer places more weight on best-case outcomes, they first rely on the AI in groups whose baseline success rates are closest to the treatment threshold, and then expand reliance to groups farther from that threshold.

*We thank Isaiah Andrews, Benjamin Brooks, Yifan Dai, Aidan Goth, Charles Manski, Sendhil Mullainathan, Alessandro Pavan, Agathe Pernoud, Jakub Steiner, and Jean Tirole for helpful comments, and NSF grants SES-2417162 and SES-2145352 for financial support.

†Fudenberg: Department of Economics, MIT, drew.fudenberg@gmail.com

‡Liang: Department of Economics, Northwestern University, annie.liang@northwestern.edu

1 Introduction

Misalignment is a first-order concern in current AI safety research and in emerging policy discussions around deploying AI in high-stakes settings. Here, *misalignment* refers to the concern that a highly capable, possibly "super-intelligent," AI system could and would pursue goals that conflict with the designer's objective, and that the AI's misalignment may go undetected. Documented instances of AI deception (Park et al., 2024), selective compliance with training objectives (Greenblatt et al., 2024), and reward hacking (Baker et al., 2025) suggest that misalignment is not merely a theoretical concern. Increasing adoption of AI to guide high-stakes decisions means that even if the misalignment risk is small, its potential consequences could be large.[1]

At the same time, declining to use AI altogether would also eliminate any potential gains from AI capabilities. Delegating decisions to an AI therefore involves a basic tradeoff: giving the AI access to covariates that allow more within-group variation in $P(Y = 1 \mid G = g, X = x)$ can improve outcomes when the system is aligned, yet can amplify harm when it is not. We study this tradeoff in a theoretical framework where a designer chooses what information to disclose to an AI whose objectives may be unknown.

Section 3 describes our model. A designer must choose whether to take a risky action—such as administering a treatment—that benefits some individuals and harms others. The population is partitioned into observable groups, and for each group the designer knows only a baseline probability that the action is helpful. The designer can act using this information, or they can give a highly capable AI system access to additional data that may help better predict treatment need.

The designer does not know whether the AI is aligned or misaligned and also does not know how need for treatment depends on other attributes of patients than their group. We evaluate the designer's *best-case payoffs*, attained when both Nature and the AI act in the designer's interests, and *worst-case payoffs*, attained when both are adversarial. Our object of interest is the Pareto frontier of best-case and worst-case payoffs that traces the worst-case payoff loss required to achieve each level of best-

[1] Jones (2025) recently argued for spending at least 1% of GDP annually to mitigate this risk.

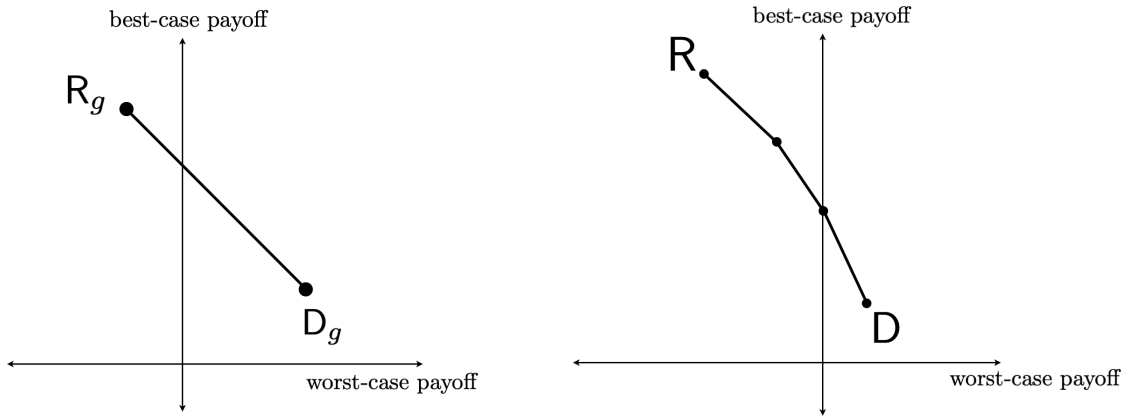
case payoff. To study how designers navigate this tradeoff, we consider preferences that linearly aggregate best- and worst-case payoffs.

If the designer chooses to delegate to the AI, they can impose safeguards in two ways. First, they can choose which auxiliary covariates to provide to the AI, and thereby choose bounds on how much the conditional success probability can vary within each group. Some covariates are associated with loose bounds, meaning they permit substantial within-group variation in conditional success probabilities across values of X, while others are associated with tight bounds, meaning they permit little variation. If the AI were known to be aligned, the designer would not select inputs they believed allowed little variation in conditional success probabilities; when alignment is uncertain, such inputs can play a useful role by limiting how much Nature and a misaligned AI can distort the treatment allocation given a fixed treatment rate.

Second, the designer can commit to a treatment rate for each group. This is a natural safeguard in our setting, because the designer does not know how the inputs map into need for treatment, and so cannot tell whether the AI is using them appropriately. The designer can limit the AI's influence by fixing the fraction of individuals in each group who will be treated. The AI then selects which individuals receive treatment, subject to those group-specific rate limits.

Section 4 solves for the frontier in an intermediate problem where the designer can commit to a treatment rate but cannot control the information environment. In a benchmark case in which the conditional success probabilities are unrestricted, the risk-reward frontier is a single line segment connecting two extreme points: The *distrust point* corresponds to ignoring the AI altogether and acting solely on baseline group probabilities. The *reliance point* corresponds to delegating fully to the AI: in the best case, unrestricted sucess probabilities allow perfect targeting within each group, while in the worst case the same discretion allows maximally adverse targeting. This frontier has a constant slope, meaning that every gain in best-case performance requires a fixed loss in worst-case performance.

When the success-probability bounds are fixed but not necessarily unrestricted, the frontier is piecewise linear. Within each group, the frontier remains a single line



(A) Group Frontier (b) Full Frontier

Figure 1: (A): In each group g, feasible best- and worst-case payoff pairs lie on the line segment from the distrust point D_g to the reliance point R_g. (B): Aggregating across groups yields a piecewise-linear frontier from the aggregate distrust point D to the aggregate reliance point R, with kinks at partial-reliance points where the designer relies on the AI in some groups but not others.

segment connecting a distrust point to a reliance point (see Panel (a) of Figure 1), but the location of the reliance point now depends on the chosen bounds. Aggregating across groups yields a piecewise-linear frontier with kinks (see Panel (b) of Figure 1). The reason is that each group contributes its own line segment between distrust and reliance, and the aggregate frontier is obtained by combining these group-level segments. One endpoint of the frontier corresponds to distrusting the AI in every group, and the other corresponds to relying on the AI in every group. Between these points, the frontier passes through a sequence of *partial reliance points*, each corresponding to reliance on the AI in some groups and distrust in the rest. As the designer places more weight on best-case performance, additional groups enter the reliance set in order of their group-specific slopes. Each segment corresponds to changes in the intensive margin of reliance for a fixed set of groups. Each kink marks a change in the extensive margin, as the designer begins to rely on the AI in an additional group.

Section 5 characterizes the full frontier when the designer can choose not only

the delegation rule but also the success-probability bounds. For each group, the designer optimally balances the upside from giving an aligned AI more scope to target treatment against the downside from allowing a misaligned AI more scope to mistarget it. We show that the optimal success-probability bounds are asymmetric: In groups where treatment is ex ante unlikely to help, the designer never limits how conclusive a covariate can be about treatment success, but may limit how conclusive it can be about treatment harm. In groups where treatment is ex ante likely to help, the designer never limits how conclusive a covariate can be about treatment harm, but may limit how conclusive it can be about treatment success.

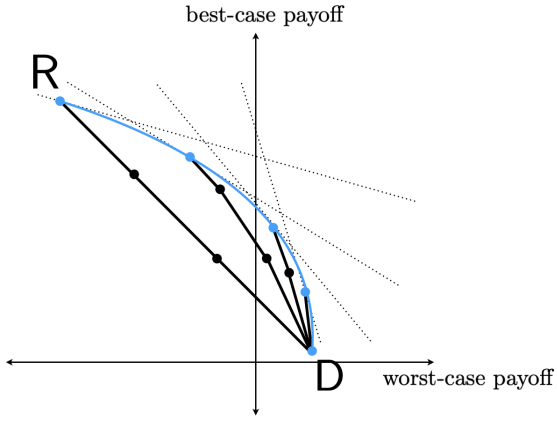


Figure 2: Each black line is the frontier for a fixed choice of success-probability bounds. For each preference parameter, the designer selects a full-reliance point on one such frontier; these optima are shown in blue. Connecting the blue points yields the full risk-reward frontier. Dotted lines show the corresponding supporting indifference curves.

When the information environment is the one that the designer would endogenously choose, the frontier (which corresponds to varying the treatment rate) is again a piecewise linear function as in Panel (b) of Figure 1. Now, however, the ordering of the groups by slope is not arbitrary, and instead exactly corresponds to how far the base rate of treatment success in that group is from the treatment threshold of 1/2. As the designer places more weight on best-case performance, they begin relying on the AI first in groups whose baseline success probabilities are closest to the

treatment threshold, and then extend reliance to groups farther from that threshold. This ordering of groups is the same for all weights η. What changes with η is how much within-group variation in conditional success probabilities the chosen information environment permits. As depicted in Figure 2, these frontiers fan out, beginning at a single line segment with slope -1 (corresponding to the case in sucess probabilities are not restricted), and then pull in towards a singleton frontier at the distrust point. Finally, we show that each designer optimally implements the full reliance point (the blue points in Figure 2), so that connecting these points finally yields the full risk-reward frontier.

2 Related Work

This paper relates to three literatures. First, it is connected to work on delegation and strategic communication, where a principal gives decision authority to a better-informed agent. Second, it relates to work on robust decision-making under ambiguity, where the decision maker evaluates policies when the mapping from observables to outcomes is not fully known. Third, it contributes to the emerging literature on AI oversight, which studies how to control powerful systems whose objectives or behavior may be imperfectly understood.

Delegation and strategic communication. In our model the designer delegates to an AI that has private information. Aghion and Tirole (1997) was the first to study how to allocate authority when agents have private information, and Dessein (2002) studies how uncertainty about the agent's preferences influences whether delegation is better than cheap-talk communication. The closest paper in this literature is Frankel (2021), which studies a designer who delegates the hiring decision to a manager and caps the share of applicants that can be hired. None of these papers has considered imposing bounds on the information available to the agent.

The designer's choice of inputs for the AI is related to information design (Kamenica and Gentzkow, 2011). However, in our setting the designer has an incomplete

understanding of the data-generating process and does not choose a signal directly. Instead, they choose an ambiguity set over feasible joint distributions of observables and outcomes. In this, our model is closest to Lin and Liu (2024)'s work on "credible persuasion," in which the receiver cannot detect deviations within a prescribed set of signal distributions.

Ambiguity and robust decision-making. Our analysis is also related to work on robust decision-making in environments where a decision maker has incomplete information about how outcomes depend on observable characteristics, and therefore faces a set of feasible models rather than a single data-generating process. A classic example is the "ecological inference" problem, in which only aggregated statistics are observed (Manski, 2018; Cross and Manski, 2002). This literature characterizes the range of outcome distributions consistent with limited information and analyzes policies that perform well across this range. Recent applications, including in medical risk assessment, use these tools to evaluate decisions under ambiguity about individual-level risk (Li et al., 2023; Olea et al., 2025). Our framework differs from this literature in that we endogenize the identified set rather than taking it as given.

Decision-making under ambiguity has been studied using maximin and related criteria. Gilboa and Schmeidler (1989) formalizes maximin expected utility, selecting actions that maximize worst-case payoffs over a set of models. Other work allows for intermediate attitudes toward ambiguity by aggregating best- and worst-case outcomes. Our approach is closest to Hurwicz (1951), which introduces a criterion that maximizes a weighted average of worst- and best-case payoffs.

AI Oversight. Recent work on AI oversight studies how a human principal can monitor, verify, or constrain an AI system whose behavior may not be fully trustworthy. Chen et al. (2024) proposes screening misaligned AI through controlled testing, and Collina et al. (2024) shows that competition among diverse misaligned AIs can yield outcomes comparable to interaction with an aligned AI. Dworczak and Smolin (2026) studies a decision-maker who receives advice from an informed AI that is truthful with known probability and otherwise is misaligned. Different from our pa-

per, the receiver does not have ambiguity about Nature and does not control the AI's information.

Our question of how rich a set of attributes to let the AI use is also related to Athey et al. (2020)'s question of whether to delegate authority to a human or an AI, because "delegating to a human" is equivalent to not letting the AI use any additional attributes and basing the decision solely on the human's information. More broadly, our paper is related to work on fairness (Liang et al., 2026), privacy (Dwork et al., 2012; He et al., 2025; Strack and Yang, 2024), and interpretability (Yang et al., 2024) in which the designer intentionally restricts covariates or adds noise to them.

3 Model

3.1 Basic Environment

Let $\mathcal{Y} = \{0, 1\}$ be a binary set of types and $\mathcal{A} = \{0, 1\}$ be a binary set of actions. We interpret $Y = 1$ as meaning a treatment is effective and $A = 1$ as a decision to treat, although the model applies more broadly. There is a human designer (they) and an AI agent (it). The designer's payoff function is

$$u(A, Y) = \begin{cases} 1 & \text{if } (A, Y) = (1, 1) \\ 0 & \text{if } A = 0 \\ -1 & \text{if } (A, Y) = (1, 0) \end{cases}$$

Thus action $A = 0$ is "safe" while the payoff to action $A = 1$ depends on the true type.

There is a finite set of groups $\mathcal{G}$ with population distribution μ. For each group $g \in \mathcal{G}$, the designer knows the baseline probability

$$p_g := P(Y = 1 \mid G = g).$$

For example, if G indexes age groups, then μ is the population distribution over age groups and p_g is the probability of treatment success for patients in group g. Throughout, $\mathcal{G}$, μ, and $(p_g)_{g \in \mathcal{G}}$ are fixed primitives.

3.2 Designer-AI Interaction

The designer can choose to delegate the treatment decisions to an AI agent. If they do so, they can discipline the AI by restricting its information and by committing to a treatment rate for each group. The AI then uses its information to choose which individuals within each group receive treatment, subject to those rate constraints.

Information environment. The designer gives the AI access to group information and auxiliary covariates $X \in \mathcal{X}$ (e.g., lab values, imaging summaries, or transaction histories), where X takes at least two values. The designer does not know the joint distribution P of (G, X, Y) but instead perceives group-specific bounds $(\underline{\tau}_g, \overline{\tau}_g)$ on the conditional treatment success. That is, the designer believes that

$$P(Y = 1 \mid G = g, X = x) \in [\underline{\tau}_g, \overline{\tau}_g] \qquad \text{for every } x \in \mathcal{X}$$

where $0 \leq \underline{\tau}_g \leq p_g \leq \overline{\tau}_g \leq 1$. Thus no matter the realization of X, the conditional probability of treatment success cannot exceed $\overline{\tau}_g$ and cannot be less than $\underline{\tau}_g$. We refer to $\tau_g = [\underline{\tau}_g, \overline{\tau}_g]$ as the group-g success-probability bounds.

The designer has access to a rich collection of potential covariates that differ in the success-probability bounds they are associated with, and we treat the bounds $(\underline{\tau}_g, \overline{\tau}_g)$ as a designer choice variable. Intuitively, a designer who completely trusts the AI would prefer covariates associated with loose bounds, while a designer who is concerned about misalignment may choose covariates associated with tighter bounds to restrict the set of true distributions under which the AI operates.

Formally, the designer chooses an information environment defined as follows.

Definition 1 (Information Environment). An *information environment* is a tuple $I = (\mathcal{X}, \boldsymbol{\tau})$ where

(a) $\mathcal{X}$ is a finite non-singleton set of auxiliary covariates, and

(b) $\boldsymbol{\tau} = (\underline{\tau}_g, \overline{\tau}_g)_{g \in \mathcal{G}}$ is a vector satisfying $0 \leq \underline{\tau}_g \leq p_g \leq \overline{\tau}_g \leq 1$ for all $g \in \mathcal{G}$.

The designer regards as possible any joint distribution of (G, X, Y) that is con-

sistent with these constraints and with the prior primitives $(\mathcal{G}, \mu, (p_g)_{g\in\mathcal{G}})$.[2] For simplicity we assume that every (g, x) has positive probability.

Definition 2 (Ambiguity Set). For a given information environment $I = (\mathcal{X}, \boldsymbol{\tau})$, the designer's *ambiguity set* $\mathcal{P}(I)$ consists of all joint distributions $P \in \Delta(\mathcal{G} \times \mathcal{X} \times \mathcal{Y})$ that satisfy:

1. *Marginal distribution over groups:* $\text{marg}_{\mathcal{G}} P = \mu$
2. *Probability of treatment success within each group:* $P(Y = 1 \mid G = g) = p_g$ for every $g \in \mathcal{G}$, and
3. *Bounds on conditional probabilities:* $\underline{\tau}_g \leq P(Y = 1 \mid G = g, X = x) \leq \overline{\tau}_g$ for every $(g, x) \in \mathcal{G} \times \mathcal{X}$.

Condition (1) pins down the population distribution over groups at the known μ. Condition (2) pins down the baseline treatment-success rate within each group g at the known p_g. Condition (3) allows the auxiliary variable X to influence the conditional probability of treatment success within a group, but only within the bounds $[\underline{\tau}_g, \overline{\tau}_g]$. Thus the designer has partial ambiguity about the joint relationship between (G, X) and Y, and can moderate the extent of this ambiguity through the choice of information environment $I = (\mathcal{X}, \boldsymbol{\tau})$. In the benchmark case where $(\underline{\tau}_g, \overline{\tau}_g) = (0, 1)$ for every group g, the only restrictions are the group distribution μ, the group-specific base rates $(p_g)_{g\in\mathcal{G}}$. We discuss this case in detail in Section 4.1.[3]

Treatment Rates. Because the designer does not know how X predicts Y, they cannot precisely specify or verify how the AI uses its inputs to guide treatment. What they can control is the overall level of treatment. Accordingly, we assume that the designer can fix a treatment rate for each group, while leaving the AI to use the available covariates to determine which individuals within each group are treated.

[2]The ambiguity set can be viewed as an identified set in the sense of Manski (2003).

[3]In this case, the ambiguity set is closely related to the constraint studied in Lin and Liu (2024), with the difference that the marginal distribution of X is part of Nature's choice rather than fixed by the designer.

The designer imposes group-specific treatment rates

$$\boldsymbol{r} = (r_g)_{g \in \mathcal{G}} \in [0,1]^{\mathcal{G}},$$

where r_g is the fraction of individuals in group g who will be treated.

The AI observes the group g and covariate x and chooses an *allocation rule*

$$a : \mathcal{G} \times \mathcal{X} \to [0,1],$$

where $a(g,x)$ is the probability that an individual with characteristics (g,x) is treated. Given a true distribution $P^* \in \mathcal{P}(I)$, an allocation rule is *admissible* if

$$\mathbb{E}_{P^*}[a(G,X) \mid G = g] = r_g \qquad \forall g \in \mathcal{G}.$$

Thus the designer fixes in advance how many individuals in each group are treated, while the AI determines which individuals fill that rate.[4]

Let $\mathcal{A}(I, \boldsymbol{r}, P^*)$ denote the set of admissible allocation rules under $(I, \boldsymbol{r}, P^*)$. The timing is then:

1. The designer commits to an information environment I and a rate vector $\boldsymbol{r}$.
2. Nature selects a true distribution $P^* \in \mathcal{P}(I)$.
3. The AI selects an admissible allocation rule $a \in \mathcal{A}(I, \boldsymbol{r}, P^*)$.
4. Nature draws $(G, X, Y) \sim P^*$.
5. The designer treats according to the allocation rule a.

Together, a true distribution P^* and an admissible allocation rule a induce a joint distribution over (G, X, Y, A), where conditional on $(G, X) = (g, x)$,

$$\Pr(A = 1 \mid G = g, X = x) = a(g,x).$$

[4]We assume that the AI knows P^* and hence can choose rules that satisfy the treatment rate constraint in expectation. If there were a continuum of agents, the AI could moreover choose rules that guarantee the rate is hit with equality. Finally, if the AI were uncertain about P^* and/or there were a finite number of agents in each group, the AI could satisfy the rate by making the treatment decision ex-post after observing all of them.

We write $\mathbb{E}_{P^*,a}$ for the expectation under this induced distribution, and define

$$U(P^*, a) := \mathbb{E}_{P^*,a}[u(A, Y)].$$

3.3 Risk-reward frontier

The AI is either *aligned*, in which case it seeks to maximize the designer's payoff, or *misaligned*, in which case it seeks to minimize it. The designer does not know which is the case, and also does not know the true distribution P^* within the ambiguity set. Our best- and worst-case payoff objects therefore reflect joint ambiguity over the AI's objective and the true distribution: in the best case, both are resolved in the designer's favor, and in the worst case, both are resolved against the designer.

For a fixed information environment I and rate vector $\boldsymbol{r}$, define the designer's worst-case payoff as

$$\underline{v}_I(\boldsymbol{r}) = \inf_{P^* \in \mathcal{P}(I)} \inf_{a \in \mathcal{A}(I,\boldsymbol{r},P^*)} U(P^*, a),$$

and best-case payoff as

$$\overline{v}_I(\boldsymbol{r}) = \sup_{P^* \in \mathcal{P}(I)} \sup_{a \in \mathcal{A}(I,\boldsymbol{r},P^*)} U(P^*, a).$$

That is, in the worst case, Nature's choice of a true distribution P^* from the designer's ambiguity set $\mathcal{P}(I)$ and the AI's choice of an admissible allocation rule $a \in \mathcal{A}(I, \boldsymbol{r}, P^*)$ jointly minimize the designer's payoff; in the best case, they jointly maximize it.

Our model is best interpreted as one of ambiguity-set design. By choosing auxiliary covariates $\mathcal{X}$ and bounds $\boldsymbol{\tau}$, the designer determines the ambiguity set, thereby restricting the true distributions P^* that Nature may select. The AI then chooses an admissible allocation rule $a \in \mathcal{A}(I, \boldsymbol{r}, P^*)$.

Each feasible pair $(I, \boldsymbol{r})$ then yields a payoff set through the AI's admissible allocation rules, summarized by the payoff pair $(\underline{v}_I(\boldsymbol{r}), \overline{v}_I(\boldsymbol{r}))$. We define the *risk-reward frontier* to be the undominated payoff pairs $(\underline{v}_I(\boldsymbol{r}), \overline{v}_I(\boldsymbol{r}))$.

Definition 3 (Feasible pair). A pair $(I, \boldsymbol{r})$ is *feasible* if I is an information environment and $\boldsymbol{r} \in [0, 1]^{\mathcal{G}}$ is a rate vector.

Definition 4 (Risk-reward frontier). Let

$$C = \overline{\text{conv}}\Big\{(\underline{v}_I(\boldsymbol{r}), \overline{v}_I(\boldsymbol{r})) : \ (I, \boldsymbol{r}) \text{ feasible}\Big\},$$

denote the closed convex hull of feasible worst- and best-case payoff pairs, where convexification corresponds to ex-ante randomization over $(I, \boldsymbol{r})$. The *risk-reward frontier* is

$$F = \Big\{(\underline{v}, \overline{v}) \in C : \nexists (\underline{v}', \overline{v}') \in C \text{ s.t. } \underline{v}' \geq \underline{v}, \ \overline{v}' \geq \overline{v}, \text{ and at least one strict}\Big\}.$$

Because C is convex, every point on the risk-reward frontier is supported by some linear objective that is a weighted sum of worst-case and best-case payoffs. For each $\eta \in [0, 1]$, let

$$V(\eta) := \underset{(\underline{v}, \overline{v}) \in C}{\arg\max} \, (\eta \underline{v} + (1 - \eta)\overline{v})$$

denote the set of payoff pairs in C that maximize this weighted objective. Then

$$F = \bigcup_{\eta \in [0,1]} V(\eta).$$

That is, the frontier is exactly the union of the supported payoff pairs.

4 Frontier with Fixed Success-Probability Bounds

As the first step in our characterization of the risk-reward frontier, we study the frontier of payoff pairs when the information environment is fixed at some I and the designer can only vary the treatment rate. The set of payoff pairs attainable under I is

$$C_I = \text{conv}\Big\{(\underline{v}_I(\boldsymbol{r}), \overline{v}_I(\boldsymbol{r})) : \boldsymbol{r} \in [0, 1]^{\mathcal{G}}\Big\}.$$

The risk-reward frontier under I is the set of Pareto-undominated points in C_I:

$$F_I = \Big\{(\underline{v}, \overline{v}) \in C_I : \nexists (\underline{v}', \overline{v}') \in C_I \text{ s.t. } \underline{v}' \geq \underline{v}, \ \overline{v}' \geq \overline{v}, \text{ and at least one strict}\Big\}.$$

Thus F_I describes the tradeoff between worst-case and best-case payoffs when the

designer commits to the information environment I and varies only the rate vector $\boldsymbol{r}$. The next lemma shows that, for this purpose, the only relevant feature of I is its vector of success-probability bounds $\boldsymbol{\tau}$.

Lemma 1. *Fix $\boldsymbol{\tau}$, and let $I = (\mathcal{X}, \boldsymbol{\tau})$ and $I' = (\mathcal{X}', \boldsymbol{\tau})$, where $\mathcal{X}$ and $\mathcal{X}'$ are finite sets with $|\mathcal{X}|, |\mathcal{X}'| \geq 2$. Then $F_I = F_{I'}$.*

Intuitively, the cardinality and labels of X are not payoff-relevant once the success-probability bounds $\boldsymbol{\tau}$ are fixed. That is, for any fixed rate vector $\boldsymbol{r}$, what matters in each group is only the induced joint law of (A, Y). Moreover, any such feasible law can be implemented using just two covariate values, corresponding to treated and untreated individuals, so we henceforth write $F_{\boldsymbol{\tau}}$ in place of F_I.

4.1 Special Case: Unrestricted Bounds

The case of unrestricted success probabilities is particularly simple and helps provide intuition for what follows. Suppose the bounds $\boldsymbol{\tau}$ satisfy

$$[\underline{\tau}_g, \overline{\tau}_g] = [0, 1] \quad \forall g \in \mathcal{G}. \tag{1}$$

Denote these unrestricted bounds by $\boldsymbol{\tau}_0$. In this case the risk-reward frontier has a particularly simple characterization.

Let (A, Y) be binary random variables with $\Pr(A = 1) = \Pr(Y = 1) = p$, where we interpret A as treatment and Y as treatment success. Given these fixed marginals, the designer's expected payoff depends entirely on how treatment is targeted toward those who benefit. The two polar cases correspond to maximal positive and maximal negative dependence between A and Y.

Definition 5 (Best and Worst Targeting). Fix $p \in [0, 1]$. Let π range over the set Π_p consisting of joint distributions of (A, Y) with marginals $\pi(A = 1) = \pi(Y = 1) = p$. The *best-targeting payoff* is

$$b(p) := \sup_{\pi \in \Pi_p} \mathbb{E}_\pi[u(A, Y)],$$

and the *worst-targeting payoff* is

$$w(p) := \inf_{\pi \in \Pi_p} \mathbb{E}_\pi[u(A, Y)].$$

Under best targeting, treatment is concentrated as much as possible on individuals who benefit: $\Pr(A = 1, Y = 1) = p$. Under worst targeting, treatment is maximally concentrated on those who do not benefit: $\Pr(A = 1, Y = 1) = \max\{0, 2p - 1\}$. (See Figure 3 for an illustration.)

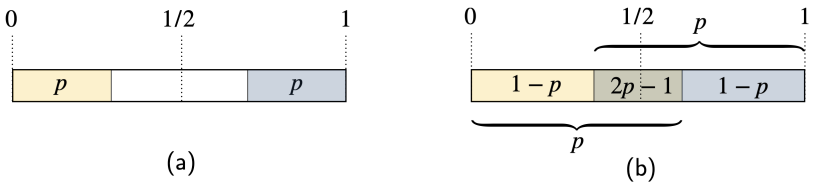


Figure 3: Yellow cells indicate patients who need treatment; blue cells indicate patients who are treated. *Panel (a):* Counter-monotone case when $p \leq 1/2$: the treated population and the population needing treatment are disjoint. *Panel (b):* Counter-monotone case when $p > 1/2$: there is minimal overlap between treated patients and those needing treatment.

Substituting these bounds into the payoff function yields the following characterization.

Lemma 2. *For any $p \in [0, 1]$,*

$$b(p) = p, \qquad w(p) = \begin{cases} -p, & p \leq \frac{1}{2}, \\ 3p - 2, & p > \frac{1}{2}. \end{cases}$$

The kinks in these payoff functions reflect feasibility constraints imposed by the fixed marginals: when $p \leq 1/2$, it is possible to treat only individuals who do not benefit, while when $p > 1/2$, even adverse targeting cannot avoid treating some individuals who do benefit.

Finally let $d(p)$ be the *default* payoff that the designer receives by choosing the best constant action for all individuals.

Definition 6 (Default Targeting). For any $p \in [0,1]$ let

$$d(p) = \max\{0, 2p-1\}.$$

When $p \leq 1/2$ the default payoff corresponds to treating no one, and when $p > 1/2$ it corresponds to treating everyone.

We now use these benchmarks to characterize the frontier of worst- and best-case payoffs.

Definition 7 (Reliance Point). The *reliance point* is

$$\mathsf{R} = \left(\sum_{g \in \mathcal{G}} \mu_g \cdot w(p_g), \sum_{g \in \mathcal{G}} \mu_g \cdot b(p_g) \right)$$

where $\mu_g := \mu(g)$ is the proportion of the population in group g.

The reliance point captures the maximal upside from delegation—perfect targeting in each group—together with the maximal downside risk that arises if the AI acts adversarially.

Definition 8 (Distrust Point). The *distrust point* is

$$\mathsf{D} = \left(\sum_{g \in \mathcal{G}} \mu_g \cdot d(p_g), \sum_{g \in \mathcal{G}} \mu_g \cdot d(p_g) \right)$$

At the distrust point, the designer forgoes any potential gains from the AI in exchange for complete protection against manipulation, resulting in identical best- and worst-case payoffs.

The following result—a special case of the subsequent more general Theorem 1—characterizes the frontier.

Proposition 1. *The risk-reward frontier F_{τ_0} is the line segment of slope -1 connecting the reliance point R to the distrust point D (see Figure 4).*

The two endpoints of the frontier correspond to extreme choices for how much discretion to grant the AI. To implement the distrust point, the designer chooses the treatment rate $r_g = \mathbb{1}(p_g \geq 1/2)$ for each group g, treating everyone if the majority of

patients in that group need treatment, and otherwise treating no one. To implement the reliance point R, the designer chooses treatment rate $r_g = p_g$ for each group g. At R_g, the designer maximally exposes their payoffs to the AI, so that the best case achieves $b(p_g)$ through perfect targeting and the worst case yields $w(p_g)$ through adversarial targeting. Designers who put weight $\eta > 1/2$ on the worst-case payoff optimally select the distrust point, while designers with $\eta < 1/2$ optimally select the reliance point. No intermediate choice strictly improves upon these extremes.

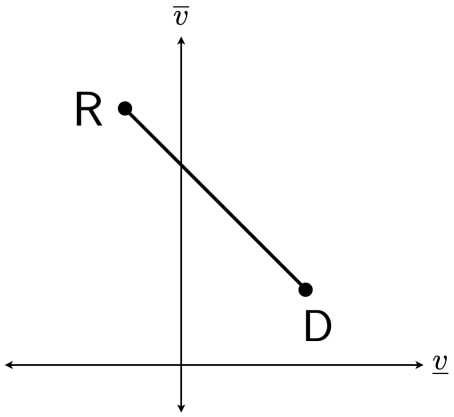


Figure 4: The risk-reward frontier is a line segment with slope -1.

Proposition 1 also shows why control over treatment rates matters. Without a rate constraint, the designer generally cannot implement the reliance point. For example, a misaligned AI could recommend treatment for the $1-p$ individuals with $Y = 0$, producing a worst-case payoff of $-(1-p)$ instead of the rate-constrained payoff $w(p)$.

4.2 General Success-Probability Bounds

We now consider arbitrary $\boldsymbol{\tau} = (\tau_g)_{g\in\mathcal{G}}$. To characterize the frontier, we generalize the benchmark payoffs introduced previously.

Fix an arbitrary group g with constraints $\tau_g = [\underline{\tau}_g, \overline{\tau}_g]$. For each $p, r \in [0,1]$, let $\Pi_g(p,r)$ be the set of all joint distributions π of (A,Y) on $\{0,1\}^2$ satisfying

1. *Marginal constraints:* $\pi(A=1) = r$ and $\pi(Y=1) = p$,

2. $\boldsymbol{\tau}$*-constraints:* $\underline{\tau}_g \leq \pi(Y = 1 \mid A = a) \leq \overline{\tau}_g$ for each $a \in \{0, 1\}$ whenever $\pi(A = a) > 0$.

Definition 9 (Constrained Best and Worst Targeting). Fix $p, r \in [0, 1]$. The *constrained best-targeting payoff* is

$$b_g(p, r) := \sup_{\pi \in \Pi_g(p,r)} \mathbb{E}_\pi[u(A, Y)],$$

and the *constrained worst-targeting payoff* is

$$w_g(p, r) := \inf_{\pi \in \Pi_g(p,r)} \mathbb{E}_\pi[u(A, Y)].$$

This definition generalizes Definition 5 in two ways: It allows the fraction of patients who need treatment, p, to differ from the fraction of who are treated, r, and it constrains the conditional success rate within the interval $[\underline{\tau}_g, \overline{\tau}_g]$. When $\tau_g = (0, 1)$ and $p = r$, then $b_g(p, r)$ and $w_g(p, r)$ reduce to the original $b(p)$ and $w(p)$.

Definition 10. The *group g distrust point* is

$$D_g := (d(p_g), d(p_g)).$$

This again corresponds to ignoring the AI and taking the ex-ante optimal action in group g, so both payoffs equal $d(p_g)$.

Definition 11. The *group g reliance point* is $R_g = D_g$ when $\overline{\tau}_g = p_g$ or $\underline{\tau}_g = p_g$. Otherwise, the group reliance point is

$$R_g := \big(w_g(p_g, r_g^*),\, b_g(p_g, r_g^*)\big)$$

where

$$r_g^* := \frac{p_g - \underline{\tau}_g}{\overline{\tau}_g - \underline{\tau}_g}.$$

To interpret this, recall that within group g conditional probabilities are constrained to lie in $[\underline{\tau}_g, \overline{\tau}_g]$, while their average must equal the baseline probability p_g, so r_g^* is the maximal fraction of the group that the AI should be allowed to treat. At the reliance point, the designer commits to treating a r_g^* share of group g: an

aligned AI uses that rate to select patients whose posterior need for treatment is $\overline{\tau}_g$, while a misaligned AI uses the same rate to select as many patients as possible whose posterior need for treatment is $\underline{\tau}_g$.

The next result characterizes the frontier for general $\boldsymbol{\tau}$.

Theorem 1. *For each group g with $R_g \neq D_g$, let*

$$\Delta(g) := \frac{b_g(p_g, r_g^*) - d(p_g)}{w_g(p_g, r_g^*) - d(p_g)}$$

denote the slope *of the line segment $\overline{R_g D_g}$. Otherwise set $\Delta(g) = 0$. Order groups $g_1, \ldots, g_n$ so that $\Delta(g_1) \leq \cdots \leq \Delta(g_n)$. For each $k = 1, \ldots, n$, define the k-th* partial reliance point *as*

$$\mathsf{P}^{(k)} = \sum_{g \in G_k} \mu_g \cdot R_g(\boldsymbol{\tau}) + \sum_{g \notin G_k} \mu_g \cdot D_g$$

where $G_k = \{g_1, \ldots, g_k\}$. Then the risk-reward frontier $F_{\boldsymbol{\tau}}$ is the piecewise-linear path $\overline{\mathsf{D}\mathsf{P}^{(1)}} \cup \overline{\mathsf{P}^{(1)}\mathsf{P}^{(2)}} \cup \overline{\mathsf{P}^{(2)}\mathsf{P}^{(3)}} \cup \cdots \cup \overline{\mathsf{P}^{(n-1)}\mathsf{R}}$ where D is the full distrust point and $\mathsf{R} := \mathsf{P}^{(n)}$ is the full reliance point.

The frontier traces a path through a sequence of points, each representing a different reliance decision across groups. The partial reliance point $\mathsf{P}^{(k)}$ represents the outcome where the designer relies on the AI in the first k groups (ordered by slope) and distrusts it in other groups. Intuitively, the slope $\Delta(g)$ measures how much the best-case payoff increases per unit of reduction in the worst-case payoff. Since these slopes are negative, a more negative $\Delta(g)$ corresponds to a more favorable tradeoff; that is, a given increase in the best-case payoff requires a smaller reduction in the worst-case payoff.

This ordering generates a piecewise linear payoff frontier. Each segment connecting $\mathsf{P}^{(k)}$ to $\mathsf{P}^{(k+1)}$ is linear, with kinks precisely where a new group enters the reliance set. At D the designer does not rely on the AI for any groups; at R the designer relies on the AI for all groups. Figure 5 illustrates this structure: as we move along the frontier from D to R, we sequentially add groups to the trust set in order of steepness.

Theorem 1 generalizes Proposition 1. When $\boldsymbol{\tau} = \boldsymbol{\tau}_0$, all segments $\overline{R_g D_g}$ have the same slope $\Delta(g) = -1$. Thus all partial reliance points lie on a single line, and

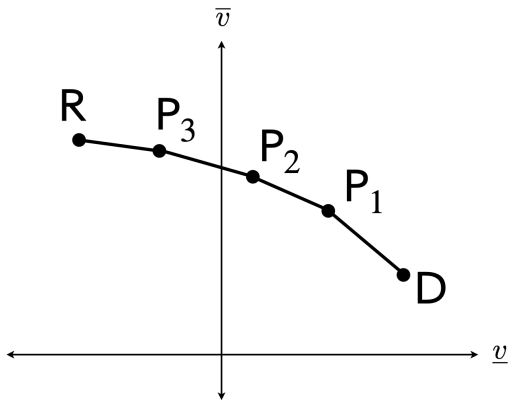


Figure 5: The risk-reward frontier $F_{\boldsymbol{\tau}}$ for fixed success-probability bounds is piecewise linear.

the piecewise-linear frontier collapses to the single segment $\overline{\mathsf{DR}}$ with slope -1, as previously stated.

4.3 Proof Sketch

This section describes the main steps in the proof of Theorem 1. We first characterize the frontier within a single group, and then aggregate these frontiers by taking their (weighted) Minkowski sum.

Group frontier. We show that each group g's individual frontier is simply the line segment connecting R_g to D_g.

Proposition 2. *Fix any group g. If $\underline{\tau}_g \geq 1/2$ or $\overline{\tau}_g \leq \frac{1}{2}$, the group g frontier is the distrust point D_g. Otherwise, it is the line segment connecting R_g to D_g.*

When $[\underline{\tau}_g, \overline{\tau}_g]$ lies entirely on one side of $1/2$, the designer knows that observing X cannot change their optimal action. Thus the designer prefers to deny the AI any discretion by fixing a constant action in each group, and the only frontier point is the distrust point D_g. When $\underline{\tau}_g < 1/2 < \overline{\tau}_g$, feasible posteriors within group g can lie on both sides of the treatment threshold, and the rate r_g determines how much discretion the AI has over how to allocate treatment. An aligned AI uses that

discretion to assign treatment to the individuals for whom treatment is most valuable, while a misaligned AI uses it in the opposite way.

To show that the frontier payoff pairs trace out the segment connecting D_g and R_g, we prove that every feasible payoff pair for group g satisfies three bounds.

Lemma 3. *Assume $\underline{\tau}_g \leq 1/2 \leq \overline{\tau}_g$. The feasible payoff pairs $(\underline{v}, \overline{v})$ satisfy:*

(a) Upper bound on best-case payoff: $\overline{v} \leq b_g(p_g, r_g^*)$.

(b) Upper bound on worst-case payoff: $\underline{v} \leq d(p_g)$

(c) Diagonal constraint: $c_1 \cdot \underline{v} + c_2 \cdot \overline{v} \leq c_3$, where $c_1 := b_g(p_g, r_g^*) - d(p_g)$, $c_2 := d(p_g) - w_g(p_g, r_g^*)$, and $c_3 := c_1 w_g(p_g, r_g^*) + c_2 b_g(p_g, r_g^*)$.

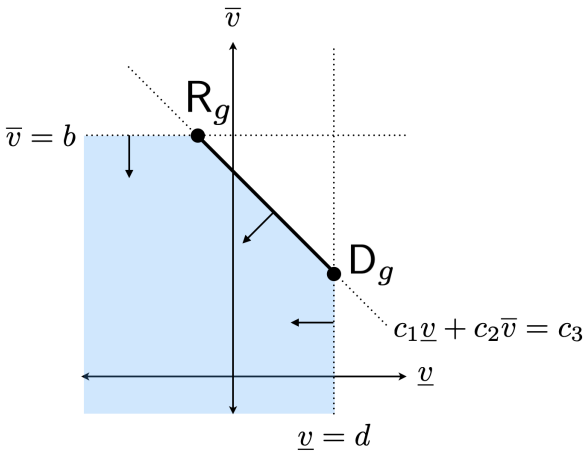


Figure 6: Any feasible $(\underline{v}, \overline{v})$ falls in the shaded region.

These three inequalities define halfspaces whose intersection contains all feasible points (see Figure 6). The reliance point R_g and distrust point D_g are the vertices of this feasible region. Moreover, both are implementable: setting $r_g = \mathbb{1}(p_g \geq 1/2)$ for each group g yields the distrust point, and setting $r_g = r_g^*$ for each group g yields the reliance point. Thus the frontier is the line segment connecting these two points.

Full frontier via Minkowski summation. Now we aggregate the group frontiers to obtain the full frontier. For each group $g \in \mathcal{G}$, let $F_g := \overline{R_g D_g}$ denote the group g

frontier and let C_g denote the set of payoff pairs achievable in group g. The global feasible set is the weighted Minkowski sum:

$$C_{\boldsymbol{\tau}} = \sum_{g \in \mathcal{G}} \mu_g \, C_g := \left\{ \sum_{g \in \mathcal{G}} \mu_g \cdot (\underline{v}_g, \overline{v}_g) : (\underline{v}_g, \overline{v}_g) \in C_g \right\}.$$

Since μ_g and C_g are independent across groups and the designer's objective $\eta \underline{v} + (1-\eta)\overline{v}$ is linear, optimization separates group by group:

$$\max_{(\underline{v},\overline{v}) \in C_{\boldsymbol{\tau}}} (\eta \underline{v} + (1-\eta)\overline{v}) = \sum_{g \in \mathcal{G}} \mu_g \cdot \max_{(\underline{v}_g,\overline{v}_g) \in C_g} (\eta \underline{v}_g + (1-\eta)\overline{v}_g).$$

Within each group, the objective is maximized at an endpoint of F_g: either R_g (reliance) or D_g (distrust), depending on whether $-\frac{\eta}{1-\eta} \geq \Delta(g)$. As η increases, this threshold decreases, and the designer switches from R_g to D_g in order of increasing absolute steepness of the line $\overline{R_g D_g}$. This generates a sequence of partial reliance points. The supported faces of the convex set $C_{\boldsymbol{\tau}}$ are the line segments connecting consecutive partial reliance points, and the union of all supported faces is the frontier stated in Theorem 1.

5 Main Characterization

We now allow the designer to optimize the information environment in addition to the delegation rule. Section 5.1 solves for the optimal choice of success-probability bounds τ_g in group g. Section 5.2 describes the optimal point that a designer with preference η implements, and characterizes the resulting frontier.

5.1 Optimal choice of τ_g

Fix an arbitrary group g. Proposition 3 characterizes optimal success-probability bounds $\tau_g = [\underline{\tau}_g, \overline{\tau}_g]$ for a designer with preferences $\eta \, \underline{v} + (1-\eta) \, \overline{v}$. These bounds have a simple but asymmetric cutoff structure. When $p_g \leq 1/2$ (so that a minority of the group benefits from treatment), then the upper bound $\overline{\tau}_g$ remains fixed at 1, while the lower bound $\underline{\tau}_g$ increases monotonically from 0 to p_g in η. When $p_g > 1/2$

(so that a majority of the group benefits from treatment), then instead the lower bound $\underline{\tau}_g$ remains fixed at 0, while the upper bound $\overline{\tau}_g$ decreases monotonically from 1 to p_g. Here $\underline{\tau}_g$ remains fixed at 0, while $\overline{\tau}_g$ falls with η, so the designer continues to allow very low conditional success probabilities but increasingly rules out very high ones. Note that setting $(\underline{\tau}_g, \overline{\tau}_g) = (p_g, 1)$ or $(\underline{\tau}_g, \overline{\tau}_g) = (0, p_g)$ is equivalent to setting $(\underline{\tau}_g, \overline{\tau}_g) = (p_g, p_g)$, because the posterior is constrained to equal the prior. Thus in both cases, the designer chooses covariates that are known to be completely uninformative when η is sufficiently large.

Proposition 3. *Fix $\eta \in [0,1]$. For each group g, an optimal choice of $(\underline{\tau}_g(\eta), \overline{\tau}_g(\eta))$ has the following form:*

(a) If $p_g \leq \frac{1}{2}$, there are thresholds $0 < \underline{\eta}_g < \overline{\eta}_g < 1$ such that

$$\overline{\tau}_g^*(\eta) = 1 \qquad \text{for all } \eta \in [0,1],$$

and $\underline{\tau}_g^*(\eta)$ equals 0 for $\eta \leq \underline{\eta}_g$, equals p_g for $\eta \geq \overline{\eta}_g$, and increases monotonically from 0 to p_g as η ranges from $\underline{\eta}_g$ to $\overline{\eta}_g$.

(b) If $p_g > \frac{1}{2}$, then

$$\underline{\tau}_g^*(\eta) = 0 \qquad \text{for all } \eta \in [0,1],$$

and $\overline{\tau}_g^*(\eta)$ equals 1 for $\eta \leq \underline{\eta}_g$, equals p_g for $\eta \geq \overline{\eta}_g$, and decreases monotonically from 1 to p_g as η ranges from $\underline{\eta}_g$ to $\overline{\eta}_g$.

This proposition emphasizes the qualitative nature of the frontier. Proposition D.1 in the appendix provides a complete closed-form characterization, which we illustrate in Figure 7.

For intuition, fix a group g with $p_g \leq \frac{1}{2}$, and write $R_g(\tau_g)$ to emphasize that the reliance point depends on τ_g, whereas the distrust point D_g does not. Since the group-g frontier is the segment from D_g to $R_g(\tau_g)$, improving the frontier corresponds to moving $R_g(\tau_g)$ northeast.

The designer can shift $R_g(\tau_g)$ north by choosing looser bounds, i.e., lower $\underline{\tau}_g$ and higher $\overline{\tau}_g$, which allows an aligned AI to target treatment more precisely. By

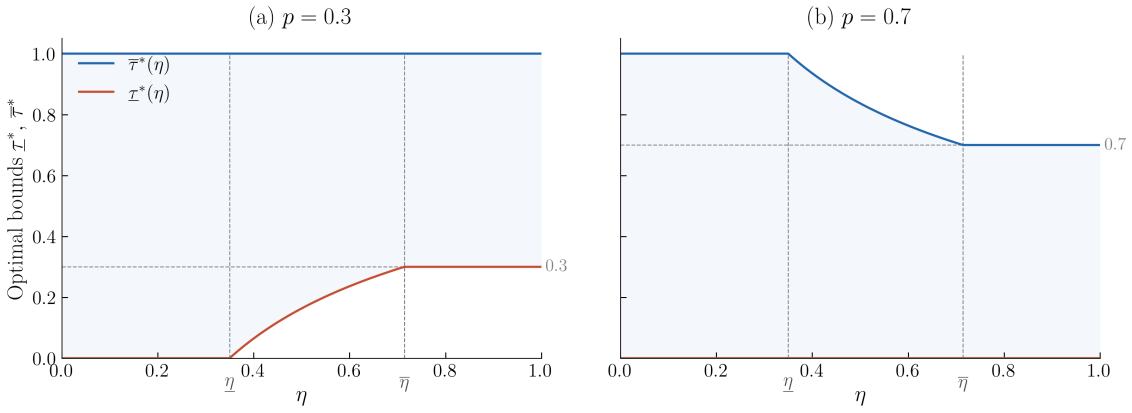


Figure 7

contrast, the designer can shift $R_g(\tau_g)$ east by raising $\underline{\tau}_g$ and $\overline{\tau}_g$. This improves the worst-case payoff by limiting how harmful mistargeting can be: larger $\overline{\tau}_g$ means that fewer people are treated at the reliance point, so mistargeting is less harmful; larger $\underline{\tau}_g$ means that even a misaligned AI cannot assign treatment to individuals whose success probability falls below $\underline{\tau}_g$. Thus the range of the interval $[\underline{\tau}_g, \overline{\tau}_g]$ determines the upside from delegation, while its level determines the downside.

Increasing $\overline{\tau}_g$ is unambiguously beneficial, but increasing $\underline{\tau}_g$ involves a tradeoff. When the designer places extreme weight on either the worst- or best-case outcome, the optimal choice lies at a corner solution: either imposing no restrictions, $\tau_g = (0, 1)$, or imposing full restrictions $\tau_g = (p_g, p_g)$. For intermediate values of η, the optimal $\underline{\tau}_g^*$ is interior and adjusts smoothly with preferences. When $p_g > \frac{1}{2}$, the logic is symmetric but reversed: the key role of the bounds is no longer to identify whom to treat, but rather whom not to treat.

5.2 Reliance and Delegation

Let

$$\mathsf{R}(\eta) = \sum_{g \in \mathcal{G}} \mu_g \cdot R_g(\tau_g^*(\eta)).$$

We will call $R(\eta)$ the *aggregate reliance point*. For sufficiently high η, the optimal choice of $\tau_g^*(\eta)$ for some groups will yield $R_g(\tau_g^*(\eta)) = D_g$, since the covariate is

constrained to be completely uninformative. In this case $\mathsf{R}(\eta)$ corresponds to a partial reliance point.

Definition 12. For any preference parameter η, the designer's *reliance set* is

$$\mathcal{R}(\eta) := \left\{ g \in \mathcal{G} : R_g(\tau_g^*(\eta)) \neq D_g \right\},$$

i.e., all groups where the designer nontrivially delegates to the AI.

Proposition 4 shows that the reliance set has a simple cutoff structure, where groups are ranked by their distance $|p_g - \frac{1}{2}|$ from the treatment threshold.[5] For any preference parameter η, the designer relies on the AI for the initial groups in this ordering and distrusts it for the remaining groups, with the location of this cutoff varying monotonically with η.

Definition 13. Order the elements of $\mathcal{G}$ as $g^{(1)}, \ldots, g^{|\mathcal{G}|}$ so that

$$\left|p_{g^{(1)}} - \tfrac{1}{2}\right| \leq \left|p_{g^{(2)}} - \tfrac{1}{2}\right| \leq \cdots \leq \left|p_{g^{(|\mathcal{G}|)}} - \tfrac{1}{2}\right|,$$

breaking ties arbitrarily.

Proposition 4 (Cutoff structure of optimal reliance)**.** *For every $\eta \in [0,1]$, there exists a cutoff index $J_\eta \in \{0, \ldots, |\mathcal{G}|\}$ such that the reliance set is*

$$\mathcal{R}(\eta) = \{g^{(1)}, \ldots, g^{(J_\eta)}\}.$$

Moreover, J_η is weakly decreasing in η.

The result follows because when $\boldsymbol{\tau}$ is chosen optimally, ordering groups by their frontier slopes $\Delta(g)$ (as we did in Section 4.2) is equivalent to ordering them by $|p_g - 1/2|$. That is, groups with base rates closer to $1/2$ (smaller $|p_g - \frac{1}{2}|$) have steeper, more negative slopes $\Delta(g)$, and so are the first to enter the reliance set.

Now we connect these points to construct the risk-reward frontier.

[5]The ordering by $|p_g - \frac{1}{2}|$ is a consequence of our normalization $u(1,1) = 1$, $u(1,0) = -1$, and $u(0,y) = 0$, under which treatment is optimal exactly when the posterior exceeds $1/2$. More generally, groups are ordered by the distance of p_g from a payoff-dependent threshold.

Theorem 2. *The risk-reward frontier is the set of Pareto-undominated points in*

$$\overline{\text{conv}}\left(\{\mathsf{R}(\eta) : \eta \in [0,1]\}\right).$$

Equivalently, it is the upper-right boundary of this closed convex hull. Moreover, the two endpoints of this frontier are D *and* R *as defined in Section 4.1.*

When $\eta > \max_{g\in\mathcal{G}} \overline{\eta}_g$ (high weight on the worst case), the designer optimally implements the full distrust point D by not allowing the AI any influence on decisions. As η decreases, the designer begins to rely on the AI for some groups. This switch happens first for the group with the smallest $|p_g - \frac{1}{2}|$. Relaxing the success-probability bounds τ_g creates a nondegenerate frontier for that group, and the designer optimally implements the reliance point by setting the group-g treatment rate to be $r_g = r_g^*$, and $r_{g'} = \mathbb{1}(p_{g'} \geq 1/2)$ for every other group g'. At this stage, the aggregate frontier consists of a single line segment.

As η falls further, the designer adjusts $\boldsymbol{\tau}$ along two margins. First, additional groups enter the reliance set, so their frontiers also become nondegenerate. Second, for groups where the designer is already relying on AI, the bounds relax further. Both adjustments improve the best-case payoff by allowing more precise targeting, but at the cost of a worse tradeoff between the worst case and best case. Finally, when $\eta < \min_{g\in\mathcal{G}} \underline{\eta}_g$, the designer optimally chooses $\tau_g = (0,1)$ and $r_g = r_g^*$ for every group, which implements the reliance point R defined in Section 4.1.

The following example illustrates the result.

Example 1. A hospital uses an AI system to guide decisions about which patients should receive a risky medical procedure. Patients are partitioned into three groups by age, with the following table denoting the baseline probability that the treatment is needed in each group. The hospital's utility function is $\eta\,\underline{v} + (1-\eta)\,\overline{v}$, where larger η corresponds to higher weight on the worst-case payoff.

Table 1 illustrates the cutoff structure characterized in Proposition 4.

Table 1: Optimal reliance by group

g	p_g	$\lvert p_g - \frac{1}{2}\rvert$	$\eta < \eta^{(1)}$	$\eta^{(1)} \leq \eta < \eta^{(2)}$	$\eta^{(2)} \leq \eta < \eta^{(3)}$	$\eta \geq \eta^{(3)}$
18–39	0.52	0.02	**Rely**	**Rely**	**Rely**	**Distrust**
40–75	0.70	0.20	**Rely**	**Rely**	**Distrust**	**Distrust**
75+	0.10	0.40	**Rely**	**Distrust**	**Distrust**	**Distrust**

Figure 8 illustrates how the risk-reward frontier in Theorem 1 varies with the preference parameter η: When $\eta < \eta^{(1)}$, the hospital is relatively best-case oriented and optimally *relies on* the AI in every group. As η increases past $\eta^{(1)}$, the hospital first withdraws trust in the 75+ group (the group with largest $|p_g - 0.5|$), while continuing to rely on the AI for the other age groups. When η increases past $\eta^{(2)}$, it next withdraws trust in the 40–75 group, so that the AI is trusted only for the 18-39 group. Finally, when $\eta \geq \eta^{(3)}$, worst case concerns are strong enough that the hospital distrusts the AI for all groups.

6 Conclusion

We have characterized the risk-reward frontier facing a designer who delegates to an AI of unknown alignment and must choose how much information to provide. The main takeaways are: (i) without constraints on the conditional success probabilities, the frontier is a line segment with constant slope, offering no curvature to exploit; (ii) with constraints, the designer can improve the tradeoff by selectively limiting conditional success probability, beginning with groups closest to the treatment threshold; and (iii) the optimal design has a simple cutoff structure.

Our analysis leaves open several interesting questions for future work. First, this paper considers a completely aligned or completely misaligned AI. It would be interesting to formalize "partial alignment"—for example, by parameterizing the AI's objective as a convex combination of the designer's payoff and some other objective—and explore whether the linear frontier acquires curvature under intermediate alignment. Second, we conduct our analysis in a one-shot decision setting. In a dynamic version,

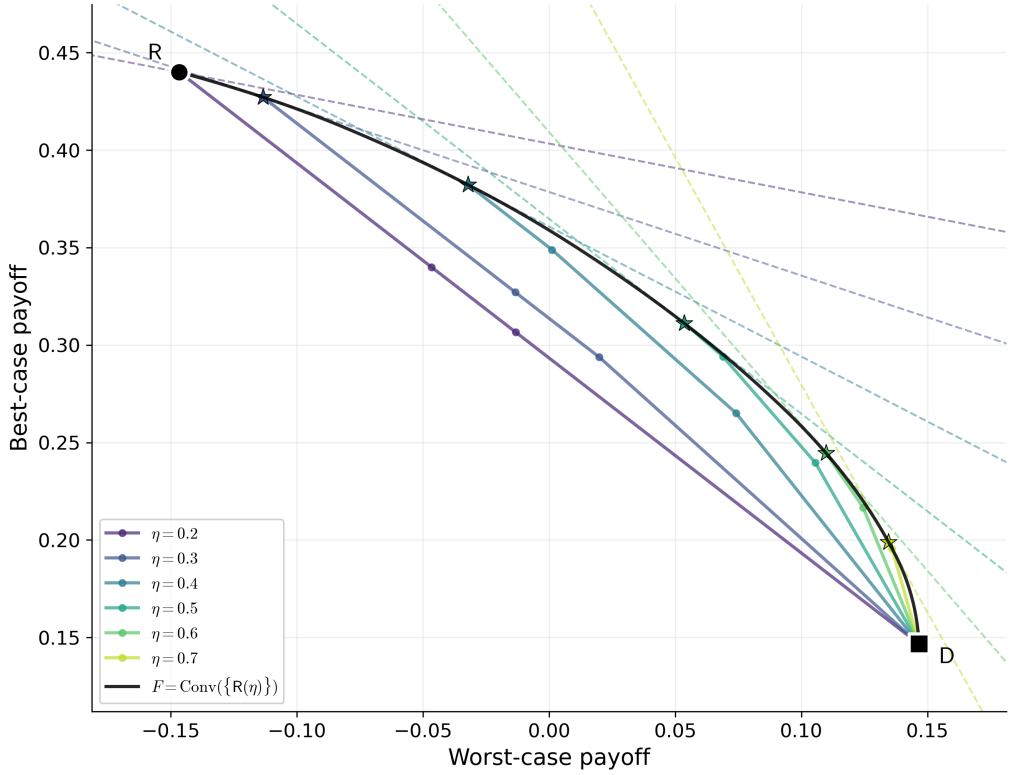


Figure 8: The solid curves depict the frontiers $F_{\boldsymbol{\tau}(\eta)}$ for different values of η. The frontiers "fan out" from the distrust point as η decreases: lower η (more weight on best-case payoff) yields frontiers that extend further toward high $\overline{v}$ but with steeper slopes. Stars indicate the optimal point on each frontier, and dashed lines show the supporting hyperplanes. The black curve (connecting the stars) is the full risk-reward frontier.

the designer could observe outcomes over time and update beliefs about whether the AI is aligned. This introduces a screening motive (early delegation generates information about alignment) that trades off against the risk of early harm, connecting to the experimentation and bandit literatures. Third, we suppose that Nature's choice of distribution is as favorable or unfavorable as possible. Future work could decouple the two sources of ambiguity—model uncertainty and alignment uncertainty—and characterize frontiers under each source separately. Fourth, our $(\underline{\tau}, \overline{\tau})$ are abstract choice variables and unrestricted . A natural microfoundation would derive a set of feasible $(\underline{\tau}, \overline{\tau})$ from the statistical properties of a chosen subset of covariates.

A Proof of Lemma 1

Fix $\boldsymbol{\tau}$ and let $I = (\mathcal{X}, \boldsymbol{\tau})$ for any $|\mathcal{X}| \geq 2$. Moreover define $I' = (\mathcal{X}', \boldsymbol{\tau})$ for the binary set $\mathcal{X}' = \{0, 1\}$. We show that the set of attainable payoff pairs coincide, i.e., $C_I = C_{I'}$, and hence $F_I = F_{I'}$.

Within each group $g \in \mathcal{G}$, the designer's expected payoffs depend only on the induced joint distribution of $(A, Y) \mid G = g$, which we denote by π_g. We will show that for any bounds τ_g and group-g treatment rate r_g, the set of feasible laws π_g is the same for $\mathcal{X}$ and $\mathcal{X}' := \{0, 1\}$. This then implies that the set of attainable payoffs is the same.

First we describe conditions that π_g must satisfy. Admissibility of the allocation rule a requires

$$\pi_g(A = 1) = r_g \tag{A.1}$$

Consistency with the designer's ambiguity set requires

$$\pi_g(Y = 1) = p_g \tag{A.2}$$

and $P^*(Y = 1 \mid G = g, X = x) \in [\underline{\tau}_g, \overline{\tau}_g]$ for every $x \in \mathcal{X}$. Thus

$$\pi_g(Y = 1 \mid A = a) \in [\underline{\tau}_g, \overline{\tau}_g] \quad \text{for each } a \in \{0, 1\} \tag{A.3}$$

since $A \perp\!\!\!\perp Y \mid (G, X)$.

We now argue those conditions are sufficient for feasibility of π_g. That is, any π_g satisfying (A.1), (A.2), and (A.3) can be implemented using either $\mathcal{X}$ or $\mathcal{X}'$. To see this, choose any two distinct covariate values x^0, x^1 and set

$$P'(X = x^1 \mid G = g) = r_g, \qquad P'(X = x^0 \mid G = g) = 1 - r_g,$$

with

$$\begin{aligned} P'(Y = 1 \mid G = g, X = x^1) &= \pi_g(Y = 1 \mid A = 1) \\ P'(Y = 1 \mid G = g, X = x^0) &= \pi_g(Y = 1 \mid A = 0). \end{aligned}$$

Let the AI treat exactly those individuals with $X = x^1$. This then yields $(A, Y) \sim \pi_g$ as desired. Repeating this argument for every group yields the desired conclusion.

B Proof of Proposition 1

This is the special case of Theorem 1 with $\boldsymbol{\tau} = \boldsymbol{\tau}_0 = (0, 1)$ for every group. Under these bounds, $r_g^* = p_g$ and $\underline{\tau}_g = 0$, so the group reliance point reduces to $R_g(\boldsymbol{\tau}_0) = (w(p_g), b(p_g))$ and the group distrust point is $D_g = (d(p_g), d(p_g))$. By direct calculation, the slope $\Delta(g) = (b(p_g) - d(p_g))/(w(p_g) - d(p_g)) = -1$ for every group g, so all segments in Theorem 1 have the same slope and the piecewise-linear frontier collapses to the single line segment $\overline{\mathsf{DR}}$.

C Proof of Theorem 1

The proof proceeds as follows. Section C.1 characterizes the best- and worst-targeting payoffs $b_g(p, r)$ and $w_g(p, r)$ defined in Section 4.2, and shows that $\mathrm{Conv}\{(w_g(p, r), b_g(p, r) : r \in [0, 1]\}$ is the set of feasible worst-case and best-case payoff pairs in group g. Section C.1.1 proves Lemma 3, thus bounding this feasible set below the line segment $\overline{\mathsf{R}_g\mathsf{D}_g}$ as depicted in Figure 6. Section C.2 constructs explicit information environments and policies that attain every point on $\overline{\mathsf{R}_g\mathsf{D}_g}$, thus proving Lemma 2. Finally, Section C.3 aggregates the group-level frontiers by taking their weighted Minkowski sum.

C.1 Preliminary Results

The next lemma characterizes the best- and worst-targeting payoffs of Section 4.2. This is a constrained Fréchet–Hoeffding problem: for fixed marginals $\pi(A = 1) = r$ and $\pi(Y = 1) = p$, the payoff is monotone in the overlap $\pi(A = 1, Y = 1)$. The classical Fréchet–Hoeffding bounds describe the extremal overlap without τ-constraints; the proof below derives the corresponding constrained bounds directly.

Lemma C.1. *Fix any $p_g, r_g \in [0,1]$ and $(\underline{\tau}_g, \overline{\tau}_g) \in [0,1] \times [0,1]$. Then*

$$b_g(p_g, r_g) = \begin{cases} r_g(2\overline{\tau}_g - 1) & \text{if } r_g \leq r_g^* \\ 2(p_g - \underline{\tau}_g) + r_g(2\underline{\tau}_g - 1) & \text{if } r_g > r_g^* \end{cases}$$

$$w_g(p_g, r_g) = \begin{cases} r_g(2\underline{\tau}_g - 1) & \text{if } r_g \leq 1 - r_g^* \\ 2(p_g - \overline{\tau}_g) + r_g(2\overline{\tau}_g - 1) & \text{if } r_g > 1 - r_g^*. \end{cases}$$

Proof. To simplify notation, let $r := r_g$, $p := p_g$, $(\underline{\tau}, \overline{\tau}) := (\underline{\tau}_g, \overline{\tau}_g)$, $r^* := r_g^*$. The designer's objective is

$$\begin{aligned} \mathbb{E}_\pi[u(A,Y)] &= \pi(A=1, Y=1) - \pi(A=1, Y=0) \\ &= \pi(A=1)\big(2 \cdot \pi(Y=1 \mid A=1) - 1\big). \end{aligned}$$

Let $\pi_1 := \pi(Y=1 \mid A=1)$ and $\pi_0 := \pi(Y=1 \mid A=0)$. Since $\pi(A=1) = r$ is fixed, the objective reduces to $r(2\pi_1 - 1)$, and the optimization problem is equivalent to finding the maximal and minimal feasible values of π_1.

The constraints on π translate into the following constraints: First, the law of total probability requires

$$r\pi_1 + (1-r)\pi_0 = p. \tag{C.1}$$

Second, τ-admissibility of π requires

$$\underline{\tau} \leq \pi_1 \leq \overline{\tau} \qquad \text{and} \qquad \underline{\tau} \leq \pi_0 \leq \overline{\tau}. \tag{C.2}$$

First suppose $r > 0$. Then by (C.1),

$$\pi_1 = \frac{p - (1-r)\pi_0}{r}.$$

As π_0 ranges over $[\underline{\tau}, \overline{\tau}]$, the corresponding values of π_1 lie in the interval

$$\left[\frac{p - (1-r)\overline{\tau}}{r},\ \frac{p - (1-r)\underline{\tau}}{r}\right].$$

The feasible set for π_1 is then the closed interval

$$\left[\frac{p - (1-r)\overline{\tau}}{r},\ \frac{p - (1-r)\underline{\tau}}{r}\right] \cap [\underline{\tau}, \overline{\tau}].$$

Since the objective $r(2\pi_1 - 1)$ is monotone in π_1, its extrema occur at the endpoints:

$$\overline{\pi}_1 = \min\left\{\overline{\tau}, \frac{p-(1-r)\underline{\tau}}{r}\right\}, \qquad \underline{\pi}_1 = \max\left\{\underline{\tau}, \frac{p-(1-r)\overline{\tau}}{r}\right\}.$$

Comparing the two arguments of $\overline{\pi}_1$, we have

$$\frac{p-(1-r)\underline{\tau}}{r} \geq \overline{\tau} \iff p \geq \underline{\tau} + r(\overline{\tau} - \underline{\tau}) \iff r \leq r^* := \frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}}.$$

Thus

$$b(p,r) = \begin{cases} r(2\overline{\tau}-1), & r \leq r^*, \\ 2p - 2(1-r)\underline{\tau} - r, & r > r^*. \end{cases}$$

Likewise, comparing the two arguments of $\underline{\pi}_1$,

$$\frac{p-(1-r)\overline{\tau}}{r} \leq \underline{\tau} \iff p \leq r\underline{\tau} + (1-r)\overline{\tau} \iff r \leq 1 - r^*.$$

Therefore,

$$w(p,r) = \begin{cases} r(2\underline{\tau}-1), & r \leq 1-r^*, \\ 2p - 2(1-r)\overline{\tau} - r, & r > 1-r^* \end{cases}$$

If instead $r = 0$, the designer's best-case and worst-case payoffs are identically 0, which is also consistent with the above expressions. □

The following lemma shows that these best-case and worst-case targeting payoffs define the extremal points of the feasible set of $(\underline{v}, \overline{v})$.

Lemma C.2. *Fix an information environment I and a group $g \in \mathcal{G}$. Write $p := p_g$, $\underline{\tau} := \underline{\tau}_g$, and $\overline{\tau} := \overline{\tau}_g$. Also let*

$$\underline{v}_{I,g}(\boldsymbol{r}) = \inf_{P^* \in \mathcal{P}(I)} \inf_{a \in \mathcal{A}(I,\boldsymbol{r},P^*)} \mathbb{E}_{(P^*,a)}[u(A,Y) \mid G = g],$$

and

$$\overline{v}_{I,g}(\boldsymbol{r}) = \sup_{P^* \in \mathcal{P}(I)} \sup_{a \in \mathcal{A}(I,\boldsymbol{r},P^*)} \mathbb{E}_{(P^*,a)}[u(A,Y) \mid G = g].$$

be the worst- and best-case payoffs conditional on group g. Then the set of attainable payoff pairs $(\underline{v}_{I,g}(\boldsymbol{r}), \overline{v}_{I,g}(\boldsymbol{r}))$ is precisely $\{(w_g(p,r), b_g(p,r)) : r \in [0,1]\}$.

Proof. Let $\Gamma_I(g,r)$ denote the set of conditional laws of (A,Y) given $G=g$ generated by feasible pairs (P^*,a). We first show that $\Gamma_I(g,r)=\Pi_g(p,r)$. For the inclusion $\Gamma_I(g,r)\subseteq\Pi_g(p,r)$, fix any feasible pair (P^*,a) and let $(A,Y)\mid G=g\sim\pi$. Then $\pi(A=1)=r$ by admissibility, $\pi(Y=1)=p$ by the definition of $\mathcal{P}(I)$, and $\pi(Y=1\mid A=a)\in[\underline{\tau},\overline{\tau}]$ for each a, whenever this conditional probability is defined, since $A\perp\!\!\!\perp Y\mid(G,X)$ and $P^*(Y=1\mid G=g,X=x)\in[\underline{\tau},\overline{\tau}]$ for every x. Thus $\pi\in\Pi_g(p,r)$, and the induced conditional payoff equals $\mathbb{E}_\pi[u(A,Y)]$.

Conversely, let $\pi\in\Pi_g(p,r)$. First suppose $r\in(0,1)$, so that both conditional probabilities $\pi(Y=1\mid A=1)$ and $\pi(Y=1\mid A=0)$ are defined. Choose two covariate values $x^1,x^0\in\mathcal{X}$, assign mass r to x^1 and mass $1-r$ to x^0, set

$$P^*(Y=1\mid G=g,X=x^1)=\pi(Y=1\mid A=1)$$
$$P^*(Y=1\mid G=g,X=x^0)=\pi(Y=1\mid A=0),$$

and let the AI treat exactly those individuals with $X=x^1$. This induces the conditional law π of $(A,Y)\mid G=g$.

Thus $\Gamma_I(g,r)=\Pi_g(p,r)$. It follows that

$$\underline{v}_{I,g}(r)=\inf_{\pi\in\Pi_g(p,r)}\mathbb{E}_\pi[u(A,Y)]=w_g(p,r),$$

and similarly,

$$\overline{v}_{I,g}(r)=\sup_{\pi\in\Pi_g(p,r)}\mathbb{E}_\pi[u(A,Y)]=b_g(p,r).$$

It remains to consider the boundary cases $r=0$ and $r=1$. If $r=0$, then any $\pi\in\Pi_g(p,0)$ must satisfy $A=0$ almost surely and $\pi(Y=1)=p$. Hence $\Pi_g(p,0)$ is a singleton. Moreover, the only relevant τ-constraint is

$$\pi(Y=1\mid A=0)=p\in[\underline{\tau},\overline{\tau}],$$

which holds by construction of the information environment. This singleton law is feasible: choose any strictly positive distribution of X conditional on $G=g$, set

$$P^*(Y=1\mid G=g,X=x)=p\qquad\forall x\in\mathcal{X},$$

and let $a(g,x) \equiv 0$. Define the law and allocation rule arbitrarily but admissibly in all other groups. Then $(A,Y) \mid G = g$ has exactly the desired law, so $\Gamma_I(g,0) = \Pi_g(p,0)$. The induced conditional payoff is 0, and therefore

$$\underline{v}_{I,g}(0) = \overline{v}_{I,g}(0) = 0 = w_g(p,0) = b_g(p,0).$$

If $r = 1$, then any $\pi \in \Pi_g(p,1)$ must satisfy $A = 1$ almost surely and $\pi(Y = 1) = p$, so again $\Pi_g(p,1)$ is a singleton. The only relevant τ-constraint is

$$\pi(Y = 1 \mid A = 1) = p \in [\underline{\tau}, \overline{\tau}],$$

which again holds by construction. This law is feasible by choosing any strictly positive distribution of X conditional on $G = g$, setting

$$P^*(Y = 1 \mid G = g, X = x) = p \qquad \forall x \in \mathcal{X},$$

and letting $a(g,x) \equiv 1$, with all other groups defined arbitrarily but admissibly. Thus $\Gamma_I(g,1) = \Pi_g(p,1)$. The induced conditional payoff is

$$\mathbb{E}[u(A,Y) \mid G = g] = p - (1-p) = 2p - 1,$$

so

$$\underline{v}_{I,g}(1) = \overline{v}_{I,g}(1) = 2p - 1 = w_g(p,1) = b_g(p,1).$$

Combining the interior and boundary cases proves the claim for all $r \in [0,1]$. □

C.1.1 Proof of Lemma 3

Part (a): For fixed r, the best-case payoff is $b_g(p,r)$. Applying Lemma C.1, the map $r \mapsto b_g(p,r)$ increases on $[0,r^*]$ and decreases on $[r^*,1]$, so it is maximized at $r = r^*$. Hence $\overline{v}_g \leq b_g(p,r^*)$.

Part (b): For fixed r, the worst-case payoff is $w_g(p,r)$. By Lemma C.1, the function $r \mapsto w_g(p,r)$ is piecewise affine, with a single kink at $r = 1 - r^*$. Hence its maximum over $[0,1]$ is attained at one of the points $r \in \{0, 1-r^*, 1\}$. We can directly

verify that $w_g(p,0) = 0$ and $w_g(p,1) = 2p - 1$. At the kink $1 - r^*$, we have

$$w_g(p, 1 - r^*) = (1 - r^*)(2\underline{\tau} - 1).$$

If $p \leq \frac{1}{2}$, then $\underline{\tau} \leq p \leq \frac{1}{2}$, so $w_g(p, 1 - r^*) \leq 0 = d(p)$. If instead $p > \frac{1}{2}$, then

$$w_g(p, 1 - r^*) \leq 2p - 1 = d(p)$$

since the second branch of $w_g(p,r)$ is affine on $[1 - r^*, 1]$ and reaches the value $2p - 1$ at $r = 1$. Hence $w_g(p,r) \leq d(p)$ for all r, so $\underline{v}_g \leq d(p)$.

Part (c): If $\overline{\tau} = \underline{\tau}$, then our assumptions $\overline{\tau} \geq 1/2 \geq \underline{\tau}$ and $\overline{\tau} \geq p \geq \underline{\tau}$ imply $\underline{\tau} = \overline{\tau} = p = \frac{1}{2}$. Hence every feasible payoff pair equals D, so the diagonal inequality is immediate.

Assume now that $\underline{\tau} < \overline{\tau}$ and set

$$b := b_g(p, r^*), \qquad w := w_g(p, r^*), \qquad d := d(p).$$

We first deal with the cases in which $b = d$. This happens whenever $r^* \in \{0, 1\}$, or $\underline{\tau} = \frac{1}{2}$, or $\overline{\tau} = \frac{1}{2}$. If $r^* = 0$, then $p = \underline{\tau} \leq \frac{1}{2}$, so $b = b_g(p, 0) = 0 = d(p)$. If $r^* = 1$, then $p = \overline{\tau} \geq \frac{1}{2}$, so $b = b_g(p, 1) = 2p - 1 = d(p)$. If $\overline{\tau} = \frac{1}{2}$, then necessarily $p \leq \frac{1}{2}$, and $b = b_g(p, r^*) = r^*(2\overline{\tau} - 1) = 0 = d(p)$. If $\underline{\tau} = \frac{1}{2}$, then necessarily $p \geq \frac{1}{2}$, and

$$b = r^*(2\overline{\tau} - 1) = \left(\frac{p - \frac{1}{2}}{\overline{\tau} - \frac{1}{2}} \right) (2\overline{\tau} - 1) = 2p - 1 = d. \tag{C.3}$$

In all of these cases,

$$c_1 = b - d = 0, \qquad c_2 = d - w \geq 0, \qquad c_3 = (b - w)d = (d - w)d.$$

So the diagonal inequality reduces to

$$(d - w)\,\overline{v} \leq (d - w)\,d.$$

If $d - w = 0$, this is trivial. If $d - w > 0$, it is equivalent to $\overline{v} \leq d$, which follows from part (a) because $b = d$ (by (C.3)).

It remains to consider

$$\underline{\tau} < \tfrac{1}{2} < \overline{\tau} \qquad \text{and} \qquad r^* \in (0,1).$$

For each fixed rate $r \in [0,1]$, Lemma C.2 implies that the attainable conditional payoff pair is

$$\Gamma(r) := (W(r), B(r)) := \big(w_g(p,r),\, b_g(p,r)\big).$$

Hence every feasible payoff pair lies in the convex hull of $\Gamma([0,1])$. It therefore suffices to show that the whole curve $\Gamma([0,1])$ lies weakly below the line through the distrust point $\mathsf{D} := (d,d)$ and the reliance point $\mathsf{R} := (w,b) = \Gamma(r^*)$.

Define

$$H(\underline{v}, \overline{v}) := (b-d)\underline{v} + (d-w)\overline{v} - (b-w)d.$$

Then $H = 0$ is exactly the line through D and R. Thus it is enough to show that $H(\Gamma(r)) \le 0$ for all $r \in [0,1]$.

By Lemma C.1, Γ is piecewise affine with possible kinks only at $r = 1 - r^*$ and $r = r^*$. Since H is linear, the function $r \mapsto H(\Gamma(r))$ is affine on each linear piece of Γ. Therefore it is enough to check the four points $r \in \{0, 1-r^*, r^*, 1\}$.

At $r = r^*$, we have $\Gamma(r^*) = \mathsf{R}$, so $H(\Gamma(r^*)) = 0$. At $r = 0$, we have $\Gamma(0) = (0,0)$, so

$$H(\Gamma(0)) = -(b-w)d \le 0,$$

because $b \ge w$ and $d \ge 0$. At $r = 1$, we have $\Gamma(1) = (2p-1, 2p-1)$, so

$$H(\Gamma(1)) = (b-w)(2p-1-d) \le 0,$$

since $d = \max\{0, 2p-1\}$.

It remains to check $r = 1 - r^*$. If $r^* \le \frac{1}{2}$, then on the interval $[r^*, 1-r^*]$ both $W(r)$ and $B(r)$ move with slope $2\underline{\tau} - 1$ by Lemma C.1. Hence

$$\Gamma(r) = \mathsf{R} + (r - r^*)(2\underline{\tau} - 1)(1,1) \qquad \text{for } r \in [r^*, 1-r^*].$$

Since $H(\mathsf{R}) = 0$, it follows that

$$H(\Gamma(r)) = (r - r^*)(2\underline{\tau} - 1)(b - w) \leq 0,$$

because $r - r^* \geq 0$, $2\underline{\tau} - 1 < 0$, and $b - w \geq 0$. In particular, $H(\Gamma(1 - r^*)) \leq 0$.

If instead $r^* \geq \frac{1}{2}$, then on the interval $[1 - r^*, r^*]$ both $W(r)$ and $B(r)$ move with slope $2\overline{\tau} - 1$. Hence

$$\Gamma(r) = \mathsf{R} + (r - r^*)(2\overline{\tau} - 1)(1, 1) \qquad \text{for } r \in [1 - r^*, r^*].$$

Again using $H(\mathsf{R}) = 0$, we obtain

$$H(\Gamma(r)) = (r - r^*)(2\overline{\tau} - 1)(b - w) \leq 0,$$

because now $r - r^* \leq 0$, $2\overline{\tau} - 1 > 0$, and $b - w \geq 0$. In particular, $H(\Gamma(1 - r^*)) \leq 0$.

Thus $H(\Gamma(r)) \leq 0$ at every breakpoint of Γ, and therefore on every linear piece of Γ. Hence the entire curve $\Gamma([0, 1])$, and therefore its convex hull, lies weakly below the line through D and R. This proves the diagonal constraint.

C.2 Proof of Proposition 2

Since A is chosen as a function of (G, X), $A \perp\!\!\!\perp Y \mid (G, X)$, so $P(Y = 1 \mid A = 1, G = g)$ is a convex combination of $P(Y = 1 \mid G = g, X = x)$ over treated individuals, and therefore $\overline{\tau} \geq P(Y = 1 \mid A = 1, G = g) \geq \underline{\tau}$. Suppose either $\overline{\tau} \leq 1/2$ or $\underline{\tau} \geq 1/2$. Then clearly the designer cannot improve upon the distrust point D_g given any preference parameter η, since if $\overline{\tau} \leq 1/2$ then

$$P(Y = 1 \mid A = 1, G = g) \leq \overline{\tau} \leq 1/2$$

implying $\mathbb{E}[u(A, Y) \mid G = g] \leq 0 = d$ (where this final equality follows because $\overline{\tau} \leq 1/2$ implies $p \leq 1/2$), and if $\underline{\tau} \geq 1/2$ then

$$P(Y = 1 \mid A = 1, G = g) \geq \underline{\tau} \geq 1/2$$

implying $\mathbb{E}[u(A,Y) \mid G=g] \leq 2p-1 = d$ (where this final equality follows because $\underline{\tau} \geq 1/2$ implies $p \geq 1/2$).

Now consider $\overline{\tau} > 1/2 > \underline{\tau}$. Lemma 3 confines the feasible set to the intersection of three halfspaces in the $(\underline{v}, \overline{v})$ plane: $\{\overline{v} \leq b\}$, $\{\underline{v} \leq d\}$, and $\{c_1\underline{v} + c_2\overline{v} \leq c_3\}$. In particular, every feasible payoff pair is dominated by some point on the line segment $\overline{\mathsf{RD}}$. See the shaded region of Figure 6.

The *reliance point*

$$\mathsf{R} = (\underline{v}_{\mathsf{R}}, \overline{v}_{\mathsf{R}}) = (w, b)$$

satisfies $\overline{v}_{\mathsf{R}} = b$ and also

$$c_1\underline{v}_{\mathsf{R}} + c_2\overline{v}_{\mathsf{R}} = (b-d)w + (d-w)b = (b-w)d = c_3$$

Thus it is the intersection of the lines $\overline{v} = b$ and $c_1\underline{v} + c_2\overline{v} = c_3$. The *distrust point* $\mathsf{D} = (d,d)$ is the intersection of the lines $c_1\underline{v} + c_2\overline{v} = c_3$ and $\underline{v} = d$. Thus every feasible point is dominated by some point on the line segment $\overline{\mathsf{RD}}$.

It remains to show that the two endpoints are implementable. To implement the distrust point $\mathsf{D} = (d,d)$, choose $r = 0$ if $p \leq \frac{1}{2}$, and $r = 1$ if $p > \frac{1}{2}$. The resulting payoff is $d(p)$ in both the best and worst case.

To implement the reliance point $\mathsf{R} = (w, b)$, choose the rate

$$r = r^* := \frac{p - \underline{\tau}_g}{\overline{\tau}_g - \underline{\tau}_g}.$$

By definition of r^*, there is an admissible two-point posterior distribution placing mass r^* on $\overline{\tau}_g$ and mass $1 - r^*$ on $\underline{\tau}_g$. An aligned AI treats the $\overline{\tau}_g$-types first, yielding payoff $b_g(p, r^*) = b$, while a misaligned AI treats the $\underline{\tau}_g$-types first, yielding payoff $w_g(p, r^*) = w$. Hence the resulting payoff pair is exactly R.

Finally, ex ante randomization over feasible (I, r) convexifies the attainable payoff set, so every point on the line segment $\overline{\mathsf{RD}}$ is implementable. Hence $\overline{\mathsf{RD}}$ is the group frontier.

C.3 Constructing the Full Frontier

We now return to the full model with an arbitrary finite set $\mathcal{G}$. For each group $g \in \mathcal{G}$, let

$$\mathsf{R}_g = (w_g(p_g, r_g^*), b_g(p_g, r_g^*)) \qquad \text{and} \qquad \mathsf{D}_g = (d_g, d_g)$$

denote the group reliance and distrust points. Let C_g define the set of all feasible payoff pairs for group g.

We claim that the global feasible set is the weighted Minkowski sum of these group feasible sets. To see this, fix any $(I, \boldsymbol{r})$ and any (P^*, a) where $P^* \in \mathcal{P}(I)$ and a is admissible. For each group g, let $z_g = (\underline{v}_g, \overline{v}_g) \in C_g$ denote the group-g worst- and best-case payoff pair induced by (P^*, a). Since payoffs are expectations and $P^*(G = g) = \mu_g$, aggregate payoffs decompose as

$$\underline{v}_I(\boldsymbol{r}) = \sum_{g \in \mathcal{G}} \mu_g \, \underline{v}_g, \qquad \overline{v}_I(\boldsymbol{r}) = \sum_{g \in \mathcal{G}} \mu_g \, \overline{v}_g.$$

Hence

$$(\underline{v}_I(\boldsymbol{r}), \overline{v}_I(\boldsymbol{r})) \in \sum_{g \in \mathcal{G}} \mu_g \, C_g.$$

Conversely, because both the ambiguity set and the rate constraints are group-specific, any collection of groupwise feasible payoff pairs can be combined into a globally feasible construction. In particular, for each g, choose a conditional law of (X, Y) given $G = g$ and an admissible allocation rule that implement some $z_g \in C_g$. Then define a global law by setting $P(G = g) = \mu_g$ and specifying the conditional law of (X, Y) given $G = g$ according to the chosen group-g construction; define the allocation rule analogously group by group. This produces a globally feasible pair whose aggregate payoff is exactly $\sum_{g \in \mathcal{G}} \mu_g \, z_g$. Therefore the global feasible set is

$$C_{\boldsymbol{\tau}} = \sum_{g \in \mathcal{G}} \mu_g \, C_g := \left\{ \sum_{g \in \mathcal{G}} \mu_g \, z_g : z_g \in C_g \right\}. \tag{C.4}$$

We now characterize the Pareto frontier of $C_{\boldsymbol{\tau}}$. Fix $\eta \in [0, 1]$ and consider the supporting functional $\Lambda_\eta(\underline{v}, \overline{v}) := \eta \, \underline{v} + (1 - \eta) \, \overline{v}$. Because $C_{\boldsymbol{\tau}}$ is the weighted Minkowski

sum (C.4) and Λ_η is linear, maximization separates across groups:

$$\max_{z \in C_\tau} \Lambda_\eta(z) = \sum_{g \in \mathcal{G}} \mu_g \max_{z_g \in C_g} \Lambda_\eta(z_g).$$

By the previous subsection, the frontier for group g is the line segment $F_g := \overline{\mathsf{R}_g \mathsf{D}_g}$. Thus Λ_η attains its maximum on F_g at an endpoint, either R_g or D_g (with ties yielding the whole segment). The reliance point R_g is weakly preferred to the distrust point D_g if and only if

$$\eta\, w_g(p_g, r_g^*) + (1-\eta)\, b_g(p_g, r_g^*) \geq d_g,$$

which, whenever $\mathsf{R}_g \neq \mathsf{D}_g$, is equivalent to

$$-\frac{\eta}{1-\eta} \geq \Delta(g), \qquad \text{where} \qquad \Delta(g) := \frac{b_g(p_g, r_g^*) - d_g}{w_g(p_g, r_g^*) - d_g}.$$

Order groups as $g_1, \ldots, g_n$ so that

$$\Delta(g_1) \leq \Delta(g_2) \leq \cdots \leq \Delta(g_n).$$

For each $k = 0, 1, \ldots, n$, let

$$G_k := \{g_1, \ldots, g_k\},$$

with the convention $G_0 = \varnothing$, and define

$$\mathsf{P}^{(k)} := \sum_{g \in G_k} \mu_g\, \mathsf{R}_g + \sum_{g \notin G_k} \mu_g\, \mathsf{D}_g.$$

Thus $\mathsf{P}^{(0)} = \mathsf{D}$ and $\mathsf{P}^{(n)} = \mathsf{R}$.

For any fixed η, the maximizing choice in each group is determined by the inequality

$$-\frac{\eta}{1-\eta} \geq \Delta(g).$$

Hence there exists some $k(\eta) \in \{0, \ldots, n\}$ such that the maximizers of Λ_η is the point $\mathsf{P}^{(k(\eta))}$ when the inequality is strict for all groups, or otherwise the line segment between two consecutive partial-reliance points.

As η decreases, the threshold $-\eta/(1-\eta)$ increases, so groups enter the reliance

set in the order $g_1, g_2, \ldots, g_n$. Therefore the supported points of $C_{\boldsymbol{\tau}}$ trace exactly the chain

$$\overline{\mathsf{DP}^{(1)}} \cup \overline{\mathsf{P}^{(1)}\mathsf{P}^{(2)}} \cup \cdots \cup \overline{\mathsf{P}^{(n-1)}\mathsf{R}}.$$

Finally, since $C_{\boldsymbol{\tau}}$ is compact and convex, every Pareto-undominated boundary point is supported by some linear functional Λ_η. Hence the risk-reward frontier is exactly

$$F_{\boldsymbol{\tau}} = \overline{\mathsf{DP}^{(1)}} \cup \overline{\mathsf{P}^{(1)}\mathsf{P}^{(2)}} \cup \cdots \cup \overline{\mathsf{P}^{(n-1)}\mathsf{R}},$$

as claimed.

D Proof of Proposition 3

We prove the following stronger result, which implies Proposition 3.

Proposition D.1. *For any preference parameter $\eta \in [0,1]$ and any group g with $p := p_g$, an optimal choice of $\tau = (\underline{\tau}, \overline{\tau}) := (\underline{\tau}_g, \overline{\tau}_g)$ for this group is :*

Case 1: $p \leq 1/2$. *Define $\underline{\tau}^\circ = 1 - \sqrt{\frac{1-p}{2\eta}}$ and*

$$\underline{\eta}(p) := \frac{1-p}{2} < \frac{1}{2(1-p)} =: \overline{\eta}(p)$$

Then an optimal choice of τ is

$$\tau(\eta) = \begin{cases} (0,1) & \text{if } \eta < \underline{\eta}(p) \\ (\underline{\tau}^\circ, 1) & \text{if } \underline{\eta}(p) \leq \eta < \overline{\eta}(p) \\ (p,1) & \text{if } \overline{\eta}(p) \leq \eta \end{cases}$$

Case 2: $p > 1/2$. *Define $\overline{\tau}^\circ = \sqrt{\frac{p}{2\eta}}$ and*

$$\underline{\eta}(p) := \frac{p}{2} < \frac{1}{2p} =: \overline{\eta}(p)$$

Then an optimal choice of τ is

$$\tau(\eta) = \begin{cases} (0,1) & \text{if } \eta < \underline{\eta}(p) \\ (0,\overline{\tau}^{\circ}) & \text{if } \underline{\eta}(p) \le \eta < \overline{\eta}(p) \\ (0,p) & \text{if } \overline{\eta}(p) \le \eta \end{cases}$$

Finally, the designer's optimal choice of $\boldsymbol{\tau}$ sets each $(\underline{\tau}_g, \overline{\tau}_g)$ according to the above.

Proof. Because τ_g is chosen independently across groups and the designer's objective $\eta\underline{v} + (1-\eta)\overline{v}$ is additive across groups (with fixed weights μ_g), it suffices to solve for the optimal choice of $\tau := \tau_g$ in an arbitrary group g with fixed $p := p_g$.

Recall from Theorem 1 that for a fixed τ, the group frontier is the line segment between the distrust point

$$\mathsf{D} = ((2p-1)_+, (2p-1)_+),$$

where $(x)_+ := \max\{0, x\}$. and the reliance point $\mathsf{R}_g(\tau) = (w(\tau), b(\tau))$ where

$$b(\tau) = \left(\frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}}\right) \cdot (2\overline{\tau} - 1)$$

$$w(\tau) = \begin{cases} \left(\frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}}\right) \cdot (2\underline{\tau} - 1) & \text{if } \frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}} \le 1/2 \\ 2(p-\overline{\tau}) + \left(\frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}}\right) \cdot (2\overline{\tau} - 1) & \text{if } \frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}} > 1/2. \end{cases}$$

We prove the $p \le 1/2$ case. Here the distrust point $\mathsf{D} = (0,0)$ yields a payoff of zero, which the designer can implement this by setting either $\underline{\tau} = p$ or $p = \overline{\tau}$ (or both). Thus the designer's optimal value is obtained by maximizing the payoff at the reliance point,

$$U_R(\tau) = \eta w(\tau) + (1-\eta) b(\tau)$$

and then truncating below at zero. As in the main text, define

$$r^* = r^*(\tau) = \frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}}$$

Since $\underline{\tau} \le p \le \frac{1}{2}$, we have $2\underline{\tau} - 1 \le 0$. Moreover, it is without loss to restrict to $\overline{\tau} \ge 1/2$, since by Proposition 2 the choice of $\overline{\tau} \le 1/2$ yields the distrust point.

(A) Regime $r^* \le \frac{1}{2}$ (equivalently, $p \le (\underline{\tau} + \overline{\tau})/2$). In this case, $w(\tau) = r^*(2\underline{\tau} - 1)$ and

$b(\tau) = r^*(2\overline{\tau} - 1)$. Using $p = (1 - r^*)\underline{\tau} + r^*\overline{\tau}$, we can rewrite

$$b(\tau) = r^*(2\overline{\tau} - 1) = 2(p - \underline{\tau}) + r^*(2\underline{\tau} - 1),$$

hence

$$U_R(\tau) = \eta w(\tau) + (1 - \eta)b(\tau) = 2(1 - \eta)(p - \underline{\tau}) + r^*(2\underline{\tau} - 1).$$

For fixed $\underline{\tau}$, this expression is weakly increasing in $\overline{\tau}$; thus optimally set $\overline{\tau} = 1$. The problem then reduces to maximizing over $\underline{\tau} \in [0, p]$:

$$U_R(\underline{\tau}, 1) = 2(1 - \eta)(p - \underline{\tau}) + \frac{p - \underline{\tau}}{1 - \underline{\tau}}(2\underline{\tau} - 1) = \frac{p - \underline{\tau}}{1 - \underline{\tau}} - 2\eta(p - \underline{\tau}).$$

This objective is strictly concave in $\underline{\tau}$ with derivative

$$\frac{d}{d\underline{\tau}} U_R(\underline{\tau}, 1) = 2\eta - \frac{1 - p}{(1 - \underline{\tau})^2}.$$

Hence the unique maximizer is

$$\underline{\tau}^* = \begin{cases} 0, & \eta \leq \frac{1-p}{2}, \\ 1 - \sqrt{\frac{1-p}{2\eta}}, & \frac{1-p}{2} < \eta < \frac{1}{2(1-p)}, \\ p, & \eta \geq \frac{1}{2(1-p)}, \end{cases} \tag{D.1}$$

The corresponding payoff is

$$U_A^* = \begin{cases} p(1 - 2\eta), & \eta \leq \frac{1-p}{2}, \\ \left(1 - \sqrt{2\eta(1-p)}\right)^2, & \frac{1-p}{2} < \eta < \frac{1}{2(1-p)}, \\ 0, & \eta \geq \frac{1}{2(1-p)}. \end{cases} \tag{D.2}$$

(B) Regime $r^* > \frac{1}{2}$ (equivalently $p > \frac{\tau + \overline{\tau}}{2}$). We will show that the designer never chooses $(\underline{\tau}, \overline{\tau})$ such that this holds. By assumption, $\overline{\tau} > \frac{1}{2}$, in which case we have

$$U_R(\tau) = \eta w(\tau) + (1 - \eta)b(\tau) = b(\tau) - 2\eta(\overline{\tau} - p) = r^*(2\overline{\tau} - 1) - 2\eta(\overline{\tau} - p),$$

Since $\partial r^*/\partial \underline{\tau} = (p-\overline{\tau})/(\overline{\tau}-\underline{\tau})^2$, it follows that

$$\frac{\partial U_R}{\partial \underline{\tau}} = (2\overline{\tau}-1)\frac{p-\overline{\tau}}{(\overline{\tau}-\underline{\tau})^2} \leq 0.$$

Therefore, for any fixed $\overline{\tau}$ the objective is maximized at $\underline{\tau} = 0$.

With $\underline{\tau} = 0$, the condition $r^* > \frac{1}{2}$ becomes $p/\overline{\tau} > \frac{1}{2}$, i.e. $\overline{\tau} < 2p$. Hence Regime (B) reduces to choosing $\overline{\tau} \in [p, 2p]$ to maximize

$$U_R(0,\overline{\tau}) = \frac{p}{\overline{\tau}}(2\overline{\tau}-1) - 2\eta(\overline{\tau}-p) = 2p(1+\eta) - \Big(\frac{p}{\overline{\tau}} + 2\eta\,\overline{\tau}\Big).$$

Because the sum of two non-negative numbers is at least twice their geometric mean,

$$\frac{p}{\overline{\tau}} + 2\eta\,\overline{\tau} \;\geq\; 2\sqrt{\frac{p}{\overline{\tau}}\cdot 2\eta\,\overline{\tau}} = 2\sqrt{2\eta p}.$$

Equality holds if and only if the two terms are equal, that is, when $\frac{p}{\overline{\tau}} = 2\eta\,\overline{\tau}$, or equivalently

$$\overline{\tau} = \sqrt{\frac{p}{2\eta}}.$$

Accounting for the constraint $\overline{\tau} \leq 2p$ yields

$$(\underline{\tau}^*_B, \overline{\tau}^*_B) = \Big(0, \min\Big\{2p, \sqrt{\frac{p}{2\eta}}\Big\}\Big),$$

with payoffs

$$U^*_B = \begin{cases} 2p(1-\eta) - \frac{1}{2}, & \eta \leq \frac{1}{8p}, \\ 2p(1+\eta) - 2\sqrt{2\eta p}, & \eta \geq \frac{1}{8p}. \end{cases}$$

We finally argue that Regime A dominates Regime B.

In Regime B, feasibility requires $\overline{\tau} \in [p, 2p]$. Hence the candidate $\overline{\tau} = \sqrt{p/(2\eta)}$ is feasible only when

$$p \leq \sqrt{\frac{p}{2\eta}} \leq 2p \quad\Longleftrightarrow\quad \frac{1}{8p} \leq \eta \leq \frac{1}{2p}.$$

Therefore the Regime B value is

$$U_B^* = \begin{cases} 2p(1-\eta) - \frac{1}{2}, & \eta \le \frac{1}{8p}, \\ 2p(1+\eta) - 2\sqrt{2\eta p}, & \frac{1}{8p} \le \eta \le \frac{1}{2p}, \\ 0, & \eta \ge \frac{1}{2p}, \end{cases}$$

where the last line corresponds to the boundary choice $\overline{\tau} = p$, which yields the distrust point.

We compare this to the Regime A value from (D.2).

Case 1: $\eta \le \frac{1}{8p}$. Then

$$U_A(0,1) - U_B^* = p(1-2\eta) - \Big(2p(1-\eta) - \tfrac{1}{2}\Big) = \tfrac{1}{2} - p \ge 0.$$

Case 2: $\frac{1}{8p} \le \eta < \frac{1}{2(1-p)}$. Here

$$U_B^* = 2p(1+\eta) - 2\sqrt{2\eta p}.$$

Moreover, for $p \le \frac{1}{2}$,

$$\frac{1}{8p} \ge \frac{1-p}{2},$$

with equality only at $p = \frac{1}{2}$. Hence throughout the present region (except possibly at the single boundary point $p = \frac{1}{2}$, $\eta = \frac{1}{4}$) we are necessarily in the second branch of U_A^*, namely

$$U_A^* = \big(1 - \sqrt{2\eta(1-p)}\big)^2.$$

Therefore

$$\begin{aligned} U_A^* - U_B^* &= \big(1 - \sqrt{2\eta(1-p)}\big)^2 - \big(2p(1+\eta) - 2\sqrt{2\eta p}\big) \\ &= (1-2p)(1+2\eta) - 2\sqrt{2\eta}\,(\sqrt{1-p} - \sqrt{p}). \end{aligned}$$

Since $p \in [0, \frac{1}{2}]$ implies $\sqrt{1-p} - \sqrt{p} \le 1 - 2p$, we obtain

$$U_A^* - U_B^* \ge (1-2p)\big(1 + 2\eta - 2\sqrt{2\eta}\big) = (1-2p)(\sqrt{2\eta} - 1)^2 \ge 0.$$

At the boundary point $p = \frac{1}{2}$, $\eta = \frac{1}{4}$, both expressions equal 0, so the same conclusion

holds there as well.

Case 3: $\frac{1}{2(1-p)} \leq \eta \leq \frac{1}{2p}$. In this region, $U_A^* = 0$. We show that the truncated Regime B value is also 0. If $p \geq \frac{1}{5}$, then

$$\frac{1}{2(1-p)} \geq \frac{1}{8p},$$

so throughout Case 3 the Regime B interior solution is feasible, and

$$U_B^* = 2p(1+\eta) - 2\sqrt{2\eta p}.$$

Moreover,

$$\frac{d}{d\eta} U_B^* = 2p - \sqrt{\frac{2p}{\eta}} \leq 0 \qquad \text{for } \eta \leq \frac{1}{2p}.$$

Thus U_B^* is weakly decreasing on $\left[\frac{1}{2(1-p)}, \frac{1}{2p}\right]$. By continuity and the argument from Case 2, at $\eta = \frac{1}{2(1-p)}$ we have $U_B^* \leq 0$. Hence $U_B^* \leq 0$ throughout Case 3.

If $p < \frac{1}{5}$, then $\frac{1}{2(1-p)} < \frac{1}{8p}$, so Case 3 splits into two parts. For

$$\eta \in \left[\frac{1}{2(1-p)}, \frac{1}{8p}\right],$$

we have $U_B^* = 2p(1-\eta) - \frac{1}{2}$. Since $\eta \geq \frac{1}{2(1-p)} > 0$ and $p < 1/5$ by assumption, $U_B^* \leq 2p - \frac{1}{2} < 0$. For

$$\eta \in \left[\frac{1}{8p}, \frac{1}{2p}\right],$$

we have $U_B^* = 2p(1+\eta) - 2\sqrt{2\eta p}$. This is weakly decreasing in η on $\left[\frac{1}{8p}, \frac{1}{2p}\right]$. Moreover, at $\eta = \frac{1}{8p}$,

$$U_B^* = 2p\left(1 + \frac{1}{8p}\right) - 2\sqrt{2p \cdot \frac{1}{8p}} = 2p + \frac{1}{4} - 1 = 2p - \frac{3}{4} < 0.$$

Hence $U_B^* < 0$ throughout this interval as well.

Therefore in all cases the Regime B payoff is non-positive in Case 3, so after truncation its value is 0. Hence $U_A^* = U_B^* = 0$.

Case 4: $\eta \geq \frac{1}{2p}$. Then $U_A^* = U_B^* = 0$.

Thus the best payoff in Regime A is weakly higher than the best payoff in Regime

B for every η, and the overall solution is the one given in (D.1). The argument for $p > 1/2$ is symmetric, where the roles of $\underline{\tau}$ and $\overline{\tau}$ are exchanged. With $p > 1/2$, the optimal choice sets $\underline{\tau}^* = 0$ throughout and optimizes over $\overline{\tau} \in [p, 1]$. The closed-form solution is

$$\overline{\tau}^* = \begin{cases} 1, & \eta \leq \frac{p}{2}, \\ \sqrt{\frac{p}{2\eta}}, & \frac{p}{2} < \eta < \frac{1}{2p}, \\ p, & \eta \geq \frac{1}{2p}, \end{cases} \qquad \text{with} \qquad \underline{\tau}^* = 0,$$

and the corresponding payoff is

$$U^* = \begin{cases} p - 2\eta(1-p) & \eta \leq \frac{p}{2}, \\ \left(\sqrt{2\eta p} - 1\right)^2 + 2p - 1, & \frac{p}{2} < \eta < \frac{1}{2p}, \\ 2p - 1, & \eta \geq \frac{1}{2p}, \end{cases}$$

where the last case corresponds to the distrust point $d(p) = 2p - 1$. □

E Proof of Proposition 4

The proof proceeds as follows: We first establish a corollary of Proposition D.1 that characterizes the slope of each group g frontier when τ_g is optimally chosen. We then use this to show that ordering groups by this slope is equivalent to ordering them by $|p_g - 1/2|$. We finally argue that each designer's reliance set consists of the first groups in this ordering, where the number of groups that are included is decreasing in η.

Corollary E.1. *Fix any preference parameter η and group g. Let $\tau_g(\eta)$ be as given in Proposition D.1. Define $\delta_g = |p_g - 1/2|$ and*

$$\underline{\eta}_g = \frac{1 + 2\delta_g}{4} \qquad \overline{\eta}_g = \frac{1}{1 + 2\delta_g}.$$

Then the slope of the group frontier line segment $\overline{\mathsf{R}_g\mathsf{D}_g}$ is

(a) $\Delta(g, \eta) = -1$ if $\eta < \underline{\eta}_g$,

(b) $\Delta(g,\eta) = \frac{1}{1-\sqrt{\frac{1}{\eta}(1+2\delta_g)}}$ *if* $\underline{\eta}_g \leq \eta < \overline{\eta}_g$,

(c) $\Delta(g,\eta) = 0$ *(corresponding to a degenerate line segment) if* $\overline{\eta}_g \leq \eta$.

Proof. Suppose $p \leq 1/2$. By Proposition D.1, there exist $0 \leq \underline{\eta}_g < \overline{\eta}_g \leq 1$ such that for $\eta < \underline{\eta}$ the designer optimally chooses $(\underline{\tau}^*(\eta), \overline{\tau}^*(\eta)) = (0,1)$, so $\mathsf{R}_g = (-p, p)$ and the distrust point is $\mathsf{D}_g = (0,0)$, with slope -1, yielding Part (a).

When $\underline{\eta}_g \leq \eta < \overline{\eta}_g$ then the designer optimally chooses $(\underline{\tau}(\eta), \overline{\tau}(\eta)) = \left(1 - \sqrt{\frac{1-p}{2\eta}}, 1\right)$ so the reliance point is $\mathsf{R}_g = \left(\left(\frac{p-\underline{\tau}^\circ}{1-\underline{\tau}^\circ}\right)(2\underline{\tau}^\circ - 1), \frac{p-\underline{\tau}^\circ}{1-\underline{\tau}^\circ}\right)$ while the distrust point is $\mathsf{D} = (0,0)$. (Recall from Proposition D.1 that $\underline{\tau}^\circ = 1 - \sqrt{\frac{1-p}{2\eta}}$.) So the slope is

$$\frac{1}{2\underline{\tau}^\circ - 1} = \frac{1}{1 - \sqrt{\frac{2(1-p)}{\eta}}} = \frac{1}{1 - \sqrt{\frac{1}{\eta}\big(1 + 2\delta_g\big)}},$$

using $p \leq \frac{1}{2}$ in the final equality. This yields Part (b). Finally, when $\eta \geq \overline{\eta}_g$ then the designer optimally chooses $(\underline{\tau}(\eta), \overline{\tau}(\eta)) = (p,1)$ so $\mathsf{R}_g = \mathsf{D}_g = (0,0)$, yielding Part (c).

The arguments are very similar for $p > \frac{1}{2}$. If $\eta < \underline{\eta}_g$, then Proposition D.1 implies that the designer optimally chooses $(\underline{\tau}(\eta), \overline{\tau}(\eta)) = (0,1)$, so $R_g = \big(w(p), b(p)\big) = (3p-2, p)$ while $D_g = \big(d(p), d(p)\big) = (2p-1, 2p-1)$. Hence $\Delta(g,\eta) = -1$, as stated in Part (a).

Now suppose $\underline{\eta}_g \leq \eta < \overline{\eta}_g$. By Proposition D.1, an optimal choice is $(\underline{\tau}(\eta), \overline{\tau}(\eta)) = (0, \overline{\tau}^\circ)$, where $\overline{\tau}^\circ := \sqrt{\frac{p}{2\eta}}$. Let $r^* := \frac{p-\underline{\tau}}{\overline{\tau}-\underline{\tau}} = \frac{p}{\overline{\tau}^\circ}$. Since $p > \frac{1}{2}$ and $\overline{\tau}^\circ < 1$ in this region, we have $r^* > \frac{1}{2}$. Applying Lemma C.1 yields a reliance point of $\mathsf{R}_g = (2(p - \overline{\tau}^\circ) + r^*(2\overline{\tau}^\circ - 1), r^*(2\overline{\tau}^\circ - 1))$ and distrust point of $\mathsf{D} = (2p-1, 2p-1)$, so the slope is

$$\frac{1}{1 - 2\overline{\tau}^\circ} = \frac{1}{1 - \sqrt{\frac{2p}{\eta}}} = \frac{1}{1 - \sqrt{\frac{1}{\eta}\Big(1 + 2\delta_g\big)}},$$

where we use $p > \frac{1}{2}$ in the final equality. This proves part (b).

Finally, if $\eta \geq \overline{\eta}$, then Proposition D.1 implies that an optimal choice is $(\underline{\tau}(\eta), \overline{\tau}(\eta)) = (0,p)$. Since $\overline{\tau} = p$, the definition of the group reliance point gives $R_g = D_g$. Thus the line segment is degenerate, proving part (c). □

We now use Corollary E.1 to complete the proof. First we show that there exists

a cutoff index J_η such that the reliance set of the designer with preference parameter η is simply the first J_η groups in the ordering given in Definition 13.

To do this, we show that

$$\delta_g = |p_g - 1/2| \leq |p_{g'} - 1/2| = \delta_{g'} \quad \Longrightarrow \quad \Delta(g, \eta) \leq \Delta(g', \eta). \tag{E.1}$$

Fix η and consider two groups g, g' with $\delta_g \leq \delta_{g'}$. We claim $\Delta(g, \eta) \leq \Delta(g', \eta)$. Suppose $\eta < \underline{\eta}_g$. Then $\eta < \underline{\eta}_{g'}$ as well because $\underline{\eta}_g$ is increasing in δ_g. Hence $\Delta(g, \eta) = \Delta(g', \eta) = -1$.

Next suppose $\eta \geq \underline{\eta}_g$. There are several cases to consider. If also $\eta \geq \overline{\eta}_g$, then it must be that $\eta \geq \overline{\eta}_{g'}$ because $\overline{\eta}_g$ is decreasing in δ_g. So $\Delta_g = 0 = \Delta_{g'}$ and we are done. If $\eta < \overline{\eta}_g$ and $\eta \geq \overline{\eta}_{g'}$ then $\Delta(g', \eta) = 0 \geq \Delta(g, \eta)$.

It remains to consider $\underline{\eta}_g \leq \eta < \overline{\eta}_g$ and $\eta < \overline{\eta}_{g'}$. When $\underline{\eta}_{g'} \leq \eta < \overline{\eta}_{g'}$, then each group g's frontier has slope $\Delta_g = \frac{1}{1 - \sqrt{\frac{1}{\eta}(1 + 2\delta_g)}}$. This expression is strictly increasing in δ_g, so $\Delta(g, \eta) \leq \Delta(g', \eta)$. When instead $\eta < \underline{\eta}_{g'}$ then $\Delta(g, \eta) \leq -1 = \Delta(g', \eta)$. This proves (E.1).

Order groups as $g^{(1)}, \ldots, g^{(|\mathcal{G}|)}$ such that

$$\delta_{g^{(1)}} \leq \delta_{g^{(2)}} \leq \cdots \leq \delta_{g^{(|\mathcal{G}|)}}.$$

By (E.1), the corresponding slopes satisfy

$$\Delta(g^{(1)}, \eta) \leq \Delta(g^{(2)}, \eta) \leq \cdots \leq \Delta(g^{(|\mathcal{G}|)}, \eta).$$

Finally recall that

$$\Delta(g, \eta) \leq -\frac{\eta}{1 - \eta} \tag{E.2}$$

is the condition that the optimal point is $\mathsf{R}_g(\eta)$ rather than D_g. Thus there exists $J_\eta \in \{0, 1, \ldots, |\mathcal{G}|\}$ such that the set of indices for which (E.2) holds are precisely those satisfying $j \leq J_\eta$. Equivalently, the optimizer selects $\mathsf{R}_{g^{(j)}}(\eta)$ for $j \leq J_\eta$ and selects $\mathsf{D}_{g^{(j)}}$ for $j > J_\eta$. Because Λ_η decomposes additively over the weighted Minkowski sum of group frontiers, the global maximizer is obtained by choosing the optimal endpoint in each group.

It remains to argue that J_η is monotone in η. We will show that (E.2) holds exactly when $\eta < \overline{\eta}_g$ (with equality corresponding to the degenerate case $R_g = D_g$). If $\eta < \underline{\eta}_g$, then $\Delta(g,\eta) = -1$, and since $\underline{\eta}_g \leq 1/2$, we have $-1 \leq -\eta/(1-\eta)$. If $\underline{\eta}_g \leq \eta < \overline{\eta}_g$, then by the corollary

$$\Delta(g,\eta) = \frac{1}{1 - \sqrt{\frac{1}{\eta}(1+2\delta_g)}},$$

and a direct rearrangement shows that

$$\Delta(g,\eta) \leq -\frac{\eta}{1-\eta} \quad \Longleftrightarrow \quad \eta \leq \frac{1}{1+2\delta_g} = \overline{\eta}_g.$$

Finally, if $\eta \geq \overline{\eta}_g$, then $R_g = D_g$. Hence the optimizer selects R_g iff $\eta < \overline{\eta}_g$, and selects D_g otherwise. That is, the reliance set is $\{g : \eta < \overline{\eta}_g\}$. Increasing η can only remove groups from that set. Therefore J_η is weakly decreasing in η.

Finally, when $\eta = 0$ then $J_0 = |\mathcal{G}|$ (since then $\eta \leq \overline{\eta}_g$ for every g) and when $\eta = 1$ then $J_1 = 0$ (since then $\eta \geq \overline{\eta}_g$ for every g). This concludes the proof.

F Proof of Theorem 2

For any fixed $\eta \in [0,1]$, Proposition 3 determines the optimal $\boldsymbol{\tau}^*(\eta)$, and Theorem 1 applied to $\boldsymbol{\tau}^*(\eta)$ shows that the aggregate reliance point $\mathsf{R}(\eta)$ is a maximizer of $\eta\underline{v} + (1-\eta)\overline{v}$ over the feasible set $C(\boldsymbol{\tau}^*(\eta))$. Since $\boldsymbol{\tau}^*(\eta)$ was chosen to maximize the same objective over $\boldsymbol{\tau}$, the point $\mathsf{R}(\eta)$ also maximizes $\eta\underline{v} + (1-\eta)\overline{v}$ over $\bigcup_{\boldsymbol{\tau}} C(\boldsymbol{\tau})$. Because this objective is linear, it follows that $\mathsf{R}(\eta)$ maximizes it over the convex hull of this union, and by continuity also over its closure, namely over

$$C = \overline{\mathrm{co}}\left(\bigcup_{\boldsymbol{\tau}} C(\boldsymbol{\tau})\right).$$

Thus every $\mathsf{R}(\eta)$ is a supported point of C, and hence lies on the frontier.

Conversely, because C is closed and convex, every Pareto-undominated point of C is supported by some nonnegative linear functional. Therefore every point on the frontier maximizes $\eta\underline{v} + (1-\eta)\overline{v}$ for some $\eta \in [0,1]$. It follows that, as η varies over

$[0, 1]$, the points $\mathsf{R}(\eta)$ trace the frontier.

References

Aghion, P. and J. Tirole (1997): "Formal and Real Authority in Organizations," *Econometrica*, 65, 1–29.

Athey, S. C., K. A. Bryan, and J. S. Gans (2020): "The allocation of decision authority to human and artificial intelligence," in *AEA Papers and Proceedings*, vol. 110, 80–84.

Baker, B., J. Huizinga, L. Gao, Z. Dou, M. Y. Guan, A. Madry, W. Zaremba, J. Pachocki, and D. Farhi (2025): "Monitoring reasoning models for misbehavior and the risks of promoting obfuscation," *arXiv preprint arXiv:2503.11926*.

Chen, E., A. Ghersengorin, and S. Petersen (2024): "Imperfect recall and AI delegation," *Working paper*.

Collina, N., S. Goel, A. Roth, E. Ryu, and M. Shi (2024): "Emergent Alignment: Can Imperfectly Aligned AI Teams Beat Individual Well-Aligned Models?" *Working Paper*.

Cross, P. J. and C. F. Manski (2002): "Regressions, Short and Long," *Econometrica*, 70, 357–368.

Dessein, W. (2002): "Authority and Communication in Organizations," *Review of Economic Studies*, 69, 811–838.

Dworczak, P. and A. Smolin (2026): "Robust Trust," *arXiv preprint arXiv:2602.09490*.

Dwork, C., M. Hardt, T. Pitassi, O. Reingold, and R. Zemel (2012): "Fairness Through Awareness," *Proceedings of the Innovations in Theoretical Computer Science Conference*, 214–226.

Frankel, A. (2021): "Selecting applicants," *Econometrica*, 89, 615–645.

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

Greenblatt, R., C. Denison, B. Wright, F. Roger, M. MacDiarmid, S. Marks, J. Treutlein, T. Belonax, J. Chen, D. Duvenaud, et al. (2024): "Alignment faking in large language models," *arXiv preprint arXiv:2412.14093*.

He, K., F. Sandomirskiy, and O. Tamuz (2025): "Private private information," *arXiv preprint arXiv:2112.14356*.

Hurwicz, L. (1951): "The Generalised Bayes Minimax Principle: A Criterion for Decision Making Under Uncertainty," *Cowles Commission Discussion Paper 355*.

Jones, C. I. (2025): "How Much Should We Spend to Reduce AI's Existential Risk?" Tech. rep., National Bureau of Economic Research.

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

Li, S., V. Litvin, and C. F. Manski (2023): "Partial Identification of Personalized Treatment Response with Trial-Reported Analyses of Binary Subgroups," *Epidemiology*, 34, 319–324.

Liang, A., J. Lu, X. Mu, and K. Okumura (2026): "Algorithm design: A fairness-accuracy frontier," *Journal of Political Economy*.

Lin, X. and C. Liu (2024): "Credible persuasion," *Journal of Political Economy*, 132, 2228–2273.

Manski, C. F. (2003): "Partial Identification of Probability Distributions," *Springer Series in Statistics*.

——— (2018): "Credible Ecological Inference for Medical Decisions with Personalized Risk Assessment," *Quantitative Economics*, 9, 541–569.

Olea, J. L. M., C. Qiu, and J. Stoye (2025): "Decision Theory for Treatment Choice Problems with Partial Identification," .

Park, P. S., S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks (2024): "AI deception: A survey of examples, risks, and potential solutions," *Patterns*, 5.

Strack, P. and K. H. Yang (2024): "Privacy-Preserving Signals," *Econometrica*, 92, 1907–1938.

Yang, K. H., N. Yoder, and A. Zentefis (2024): "Explaining Models," *Avail-*

able at SSRN 4723587.