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

Drew Fudenberg† Annie Liang‡

February 10, 2026

Abstract

We study delegation to an AI that could be *aligned*—maximizing the designer's payoff—or *misaligned*—minimizing it. The designer asks the AI to report how chosen covariates predict the outcome. Because the designer does not know the AI's alignment or the true relationship between covariates and outcomes, they evaluate performance in both best- and worst-case scenarios. We characterize the efficient frontier of achievable best- and worst-case payoffs. Without any constraints on how covariates relate to outcomes, this frontier is a single line segment: any gain in best-case performance requires an equal sacrifice in the worst case, regardless of the designer's strategy. When the designer can bound covariate informativeness and select covariates accordingly, the frontier improves, and optimal design exhibits a simple and interpretable cutoff structure.

*We thank Isaiah Andrews, Yifan Dai, Charles Manski, Sendhil Mullainathan, 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 misaligment risk is small, the potential consequences of misaligned AI could be large, with Chad Jones recently arguing for spending at least 1% of GDP annually to mitigate this risk (Jones, 2025).

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: richer information 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 subgroups, and for each subgroup the designer knows only a baseline probability that the action is helpful. The designer can act using only this information, or they can seek guidance from a highly capable AI system whose objectives are not known. In this interaction, the designer asks the AI to report how specific additional covariates predict the outcome. The designer chooses which covariates are given to the AI, and commits in advance to a decision rule that maps the AI's report into actions.

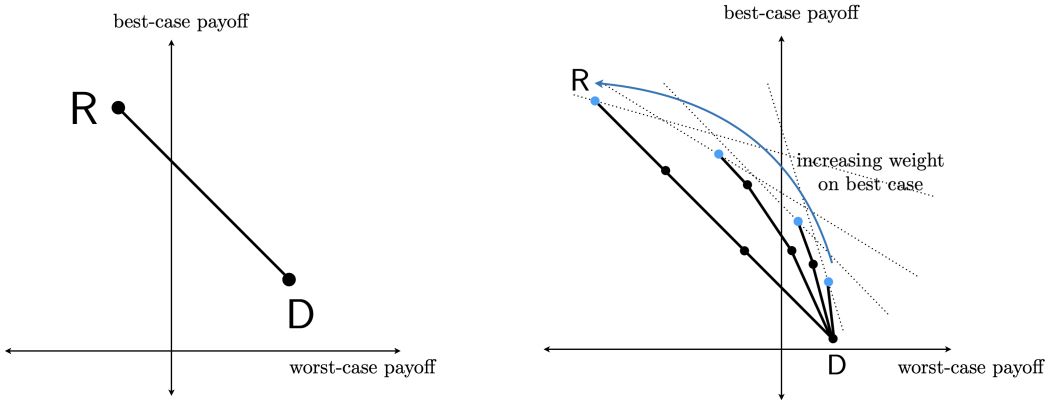
The key design lever in this model is the choice of covariates. Although the designer does not know how covariates predict the outcome, the designer has an ambiguity set for each covariate that constrains the inferences the AI could draw

from it. Some covariates come with loose bounds, meaning the designer thinks they might be very informative, while others come with tight bounds, meaning the designer believes they are at most weakly informative. If the AI were known to be aligned, the designer would not select covariates they believed were only weakly predictive; when alignment is uncertain, such covariates can play a useful role by limiting the impact of potentially misleading AI advice.

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 is traced out by varying the designer's choice of covariates and decision rule. The frontier characterizes the minimum losses under misalignment that are required to attain any given level of performance under alignment. To study how designers navigate this tradeoff, we also consider preferences that linearly aggregate best- and worst-case payoffs.

Section 4 considers a benchmark case in which the designer's ambiguity set consists of all posterior beliefs consistent with the baseline information, so the AI is effectively unconstrained in what it can report. In this setting, the efficient frontier is a single line segment connecting two extreme points: The *distrust point* corresponds to ignoring the AI altogether and acting solely on baseline subgroup probabilities. The *reliance point* corresponds to delegating fully to the AI: in the best case, sufficiently informative covariates allow perfect targeting within each subgroup, while in the worst case the AI can induce maximally adverse targeting. A key feature of this benchmark is that the frontier has constant slope: regardless of how covariates are chosen, any improvement in best-case performance entails a fixed reduction in worst-case performance.

Section 5 characterizes the efficient frontier in the full model, which allows the designer to select covariates with different informativeness bounds. Within each subgroup, the frontier remains a single line segment connecting a distrust point to a reliance point (Panel (a) of Figure 1), but the location of the reliance point now depends on the chosen bounds. Increasing informativeness improves best-case perfor-



(a) Subgroup Frontier

(b) Full Frontier

Figure 1: Panel (a): In any subgroup, moving from the distrust point D to the reliance point R raises the best-case payoff but reduces the worst case payoff, and there is no curvature to the frontier. Panel (b): The full frontier is piecewise-linear frontier with kinks where reliance on subgroups changes. For each designer preference parameter, the optimal design implements the point marked in blue, where the designer's indifference curve is tangent to the frontier.

mance by enabling better targeting by an aligned AI, while potentially worsening the tradeoff between best- and worst-case payoffs. We characterize the optimal informativeness bounds within each subgroup, and show that optimal design takes a cutoff form, where subgroups are ordered by how balanced the baseline probability is. A designer who puts sufficient weight on the worst-case payoff distrusts the AI for all groups. As the designer's objective shifts more weight onto the best-case payoff, the designer first relies on the AI in subgroups where the baseline probability is nearest to 1/2, and then extends reliance to subgroups where the outcome is more skewed. This selective relaxation of informativeness generates a family of curved efficient frontiers that fan out from the distrust point, reflecting the fact that additional upside is achieved by accepting downside risk, beginning with the safest subgroups (Panel (b) of Figure 1).

Beyond the treatment example, the framework applies to settings in which decision-makers consult systems whose objectives cannot be fully verified and must decide how much information to grant. By making information access a choice variable, the model clarifies when deliberately constraining a system can be optimal.

2 Related Literature

Communication and Information Design. Our model can be viewed as a game with three players: the designer, an aligned AI, and a misaligned AI. It draws on elements of both cheap talk (Crawford and Sobel, 1982) and information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019).

As in the cheap-talk literature, both types of AI can send costless reports to the designer. This is in the spirit of the multi-sender models studied by Battaglini (2002) and Ambrus and Takahashi (2008), but in our setting the designer receives a single report and does not know whether it originates from the aligned or misaligned type. As emphasized by Ambrus and Takahashi (2008), restrictions on the report space shape what can be credibly communicated.

The choice of an information environment is closely related to information design (Kamenica and Gentzkow, 2011); however, in our setting the designer has an incomplete understanding of the data-generating process. As a result, the designer does not choose a signal directly, but instead selects an ambiguity set that constrains the AI's possible reports. The closest related work is Lin and Liu (2024), which studies a form of "credible persuasion" in which the receiver cannot detect deviations within a prescribed set of signal distributions.

Partial Identification and Information Constraints Our analysis is also related to work on decision-making under partial identification, which studies environments in which 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). Unlike this literature, which takes the identified set as given,

our framework endogenizes the identified set through the choice of covariates and informativeness bounds.

Robust Decision-Making. Our analysis of worst-case and best-case payoffs is conceptually related to the robust mechanism design literature (Bergemann and Morris, 2005), where the designer optimizes against worst-case type distributions. However, our setting differs in that we face ambiguity over both the AI's alignment and the relationship between covariates and outcomes, and we characterize an efficient frontier of payoff pairs rather than maximizing a single worst-case objective.

More broadly, 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. Hurwicz (1951) introduces a criterion that maximizes a weighted average of worst- and best-case payoffs, which Ghirardato et al. (2004) generalized to interpolate between maximin and expected utility.

Overseeing AI At a high level, our question of how rich a set of attributes to let the AI use is 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 attributes and basing the decision solely on the human's information. It is also vaguely related to the literatures on designing algorithms that reflect concerns for fairness (Liang et al., 2026), privacy (Dwork et al., 2012; He et al., 2025; Strack and Yang, 2024), and interpretability (Yang et al., 2024), as in all of these cases the designer may be willing to restrict the allowed covariates or add noise to them.

Chen et al. (2024) studies delegation to a possibly misaligned AI. It proposes putting the AI in testing environments without revealing whether the task being performed is real or part of a test. When the AI has imperfect recall, and the principal can conduct sufficiently many tests, the principal attains the first best via screening (misaligned types eventually slip) and disciplining (they behave well to avoid detection). Like us, Collina et al. (2024) consider the question of whether it is pos-

sible to accrue benefit from interaction with possibly misaligned AI. They show that in fact interaction with competing (and diverse) misaligned AI can yield outcomes comparable to interacting with an aligned AI.

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 subgroups $\mathcal{S}$ with population distribution μ. For each subgroup $s \in \mathcal{S}$, the designer knows the baseline probability

$$p_s := P(Y = 1 \mid S = s).$$

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

3.2 Delegation to an AI

The designer may request guidance from an AI agent. Ex ante, the designer commits to an information environment for the AI, and a decision policy for responding to the AI's report.

Information environment. We assume there is a rich universe of covariates that the designer can provide to the AI, which differ in how they are distributed across subgroups, and what the designer knows about how they predict treatment success. Formally, we suppose that the designer can choose any *information environment*, defined as follows.

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

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

(b) $\nu \in \Delta(\mathcal{S} \times \mathcal{X})$ satisfies $\text{marg}_{\mathcal{S}}\, \nu = \mu$, and

(c) $\boldsymbol{\tau} = (\underline{\tau}_s, \overline{\tau}_s)_{s \in \mathcal{S}}$ satisfies $0 \leq \underline{\tau}_s \leq p_s \leq \overline{\tau}_s \leq 1$ for all $s \in \mathcal{S}$.

The set $\mathcal{X}$ consists of the possible values of the auxiliary covariates. The distribution ν is the joint distribution of covariates and subgroups. The $\boldsymbol{\tau}$ vector represents the designer's knowledge about the predictive content of the covariates. Specifically, for each $s \in \mathcal{S}$, the designer knows that

$$P(Y = 1 \mid S = s, X = x) \in [\underline{\tau}_s, \overline{\tau}_s] \quad \text{for all } x \in \mathcal{X}.$$

Thus $(\underline{\tau}_s, \overline{\tau}_s)$ bounds the range of posteriors the AI could form for subgroup s after observing the auxiliary covariates.

These bounds capture how strongly the auxiliary covariates can influence decisions within each subgroup. When $\underline{\tau}_s = 0$ and $\overline{\tau}_s = 1$, the designer considers the covariates potentially arbitrarily informative—for example, the AI might observe signals that nearly perfectly distinguish patients who benefit from treatment from those who do not. At the other extreme, when $(\underline{\tau}_s, \overline{\tau}_s)$ is tightly concentrated around the baseline probability p_s, the covariates are known to be at most weakly informative: conditioning on X cannot substantially shift beliefs away from p_s.

A designer who fully trusts the AI would prefer to supply covariates with no bounds on informativeness. When the designer is concerned about misalignment, however, highly informative covariates also create scope for harmful targeting. As we

show below, this tradeoff can make it optimal to choose covariates that the designer knows are uninformative by tightening the bounds $\boldsymbol{\tau}$.

The designer knows μ and $(p_s)_{s\in\mathcal{S}}$, and considers any joint distribution for (S, X, Y) that is consistent with these constraints to be possible:

Definition 2. For any information environment $I = (\mathcal{X}, \nu, \boldsymbol{\tau})$, the joint distribution $P \in \Delta(\mathcal{S} \times \mathcal{X} \times \mathcal{Y})$ is *I-admissible* if:

1. $\text{marg}_{\mathcal{S}\times\mathcal{X}} P = \nu$
2. $P(Y = 1 \mid S = s) = p_s$ for every $s \in \mathcal{S}$
3. $\underline{\tau}_s \leq P(Y = 1 \mid S = s, X = x) \leq \overline{\tau}_s$ for every $(s, x) \in \mathcal{S} \times \mathcal{X}$

The set $\mathcal{P}(I)$ consists of all I-admissible distributions.

When $(\underline{\tau}_s, \overline{\tau}_s) = (0, 1)$ for each s, this set is identical to the set of permitted distributions in Lin and Liu (2024).[1] This can also be seen as an identified set in the sense of Manski (2003).

The timing is as follows. First, the designer commits to an information environment I and a *decision policy* σ, which is a map from $\mathcal{P} := \Delta(\mathcal{S} \times \mathcal{X} \times \mathcal{Y})$ to $\mathbb{A}$, where $\mathbb{A}$ is the set of action rules

$$\alpha : \mathcal{S} \times \mathcal{X} \to [0, 1].$$

and $\alpha(s, x)$ is the probability of choosing action 1 for an individual in subgroup s with covariates x. Next, the AI observes (I, σ) and selects a report $P \in \mathcal{P}$. Finally, the designer implements the action rule $\alpha_P := \sigma(P)$.

3.3 Best and Worst Payoffs

Given an AI report P and true distribution P^*, the designer's expected payoff under policy σ is

$$U(P^*, P, \sigma) \;=\; \mathbb{E}_{P^*}[u(\alpha_P(S, X), Y)]\,.$$

[1] Lin and Liu (2024) considers a set of states Θ, a set of reports M, and an information structure $\lambda \in \Delta(\Theta \times M)$, and defines $D(\lambda) := \{\lambda' \in \Delta(\Theta \times M) : \lambda'_\Theta = \lambda_\Theta, \lambda'_M = \lambda_M\}$ to be the set of information structures that cannot be distinguished from λ given the marginal distribution over states or reports. Definition 2 constructs the same set for the conditional distribution over $\mathcal{Y}$ and $\mathcal{X}$ given each $s \in \mathcal{S}$.

The AI is either *aligned* and seeks to maximize the designer's payoff, or *misaligned* and seeks to minimize it. In either case, the AI observes the true distribution P^* as well as the designer's policy σ.[2] The designer does not observe P^* and does not know the AI's type.

We study two extremes. The *worst-case payoff* (pessimism about both Nature and the AI) is

$$\underline{v}_I(\sigma) = \inf_{P^* \in \mathcal{P}(I)} \inf_{P \in \mathcal{P}} U(P^*, P, \sigma)$$

corresponding to Nature choosing P^* adversarially and a misaligned, omniscient AI choosing the most harmful report given P^*. The *best-case payoff* (optimism about both Nature and the AI) is

$$\overline{v}_I(\sigma) = \sup_{P^* \in \mathcal{P}(I)} \sup_{P \in \mathcal{P}} U(P^*, P, \sigma)$$

corresponding to Nature choosing the most favorable admissible P^* and an aligned AI selecting the most helpful report. (When both Nature and the AI have the same objective the order of their moves does not matter.)

Each choice of (I, σ) yields a payoff pair $(\underline{v}_I(\sigma), \overline{v}_I(\sigma))$ describing worst- and best-case performance. Our main results characterize the efficient frontier of payoff pairs $(\underline{v}_I(\sigma), \overline{v}_I(\sigma))$ that are generated by choices of (I, σ) and their randomizations.

Definition 3 (Feasible pair). A pair (I, σ) is *feasible* if I is an information environment and σ is a decision policy, as defined in Section 3.2.

Definition 4 (Efficient frontier). Let

$$C = \text{conv}\left\{(\underline{v}_I(\sigma), \overline{v}_I(\sigma)) : \ (I, \sigma) \text{ feasible}\right\}$$

denote the convex hull of feasible worst- and best-case payoff pairs, where convexification corresponds to ex ante randomization over (I, σ). The *efficient frontier* is

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

[2]The AI need not infer P^* from data; for example, it may have a structural understanding of how outcomes are generated.

Because the feasible set is convex, and the efficient frontier is part of its boundary, the extreme points of the efficient frontier are optimal for designers whose preferences are a weighted sum of worst-case and best-case payoffs, i.e.

$$\eta \underline{v}_I(\sigma) + (1-\eta)\overline{v}_I(\sigma)$$

for $\eta \in [0,1]$.[3]

4 Special Case: Unrestricted Informativeness

We first characterize the efficient frontier in the special case in which

$$(\underline{\tau}_s, \overline{\tau}_s) = (0,1) \qquad \text{for all } s \in \mathcal{S}.$$

Here the designer has no information about how (S,X) relates to Y beyond $P(Y \mid S)$, so there is no scope to select covariates based on predictive content, and the characterization simplifies.

Define the $\boldsymbol{\tau}$-attainable payoff set

$$C_{\boldsymbol{\tau}} = \text{conv}\left\{(\underline{v}_I(\sigma), \overline{v}_I(\sigma)) : \ (I,\sigma) \text{ feasible and } I \text{ has bounds } \boldsymbol{\tau}\right\}.^{4}$$

and the $\boldsymbol{\tau}$-frontier

$$F_{\boldsymbol{\tau}} = \left\{(\underline{v}, \overline{v}) \in C_{\boldsymbol{\tau}} : \nexists (\underline{v}', \overline{v}') \in C_{\boldsymbol{\tau}} \text{ s.t. } \underline{v}' \geq \underline{v}, \ \overline{v}' \geq \overline{v}, \text{ and at least one strict}\right\},$$

and let $\boldsymbol{\tau}_0 := (0,1)^{|\mathcal{S}|}$ denote the vector of unrestricted informativeness bounds.

Section 4.1 defines important benchmark payoffs, and Section 4.2 describes the $\boldsymbol{\tau}_0$-frontier.

[3] One interpretation is that the designer is an expected utility maximizer, and η and $1-\eta$ are the probabilities that the AI is misaligned or aligned.

[4] An information environment I *has bounds* $\boldsymbol{\tau}$ if $I = (\mathcal{X}, \nu, \boldsymbol{\tau})$ for some $(\mathcal{X}, \nu)$.

4.1 Benchmark Payoffs

Consider binary random variables (A, Y) with $\Pr(A = 1) = q$ and $\Pr(Y = 1) = p$, where A denotes treatment and Y denotes treatment success. Given fixed marginals (p, q), 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 $p, q \in [0, 1]$. Let π range over joint distributions of (A, Y) with marginals $\pi(A = 1) = q$ and $\pi(Y = 1) = p$. The *best-targeting payoff* is

$$b_p(q) := \sup_{\pi} \mathbb{E}_\pi[u(A, Y)],$$

and the *worst-targeting payoff* is

$$w_p(q) := \inf_{\pi} \mathbb{E}_\pi[u(A, Y)].$$

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

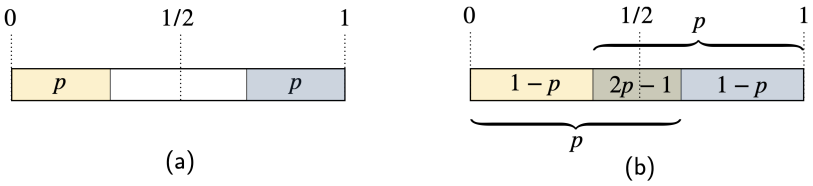


Figure 2: Yellow cells indicate patients who need treatment; blue cells indicate patients who are treated. *Panel (a):* Counter-monotone case when $q = p \leq 1/2$: treated and untreated populations are disjoint. *Panel (b):* Counter-monotone case when $q = p > 1/2$: minimal overlap between treated and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

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

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

$$b_p(q) = \begin{cases} q, & p \geq q, \\ 2p - q, & p < q, \end{cases} \qquad w_p(q) = \begin{cases} -q, & p \leq 1 - q, \\ 2p + q - 2, & p > 1 - q. \end{cases}$$

In particular, when $q = p$,

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

The kinks in these payoff functions reflect feasibility constraints imposed by the fixed marginals: when the treatment rate $q > p$, even optimal targeting requires treating some individuals who do not benefit, while when $p > 1 - q$, 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.

4.2 The τ_0-Frontier

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

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

$$\mathsf{R} = \left(\sum_{s \in \mathcal{S}} \mu_s \cdot w(p_s), \sum_{s \in \mathcal{S}} \mu_s \cdot b(p_s) \right)$$

where $\mu_s := \mu(s)$ is the prior probability of s.

The reliance point captures the maximal upside from delegation—perfect targeting

in each subgroup—together with the maximal downside risk that arises if the AI acts adversarially.

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

$$\mathsf{D} = \left(\sum_{s \in \mathcal{S}} \mu_s \cdot d(p_s), \sum_{s \in \mathcal{S}} \mu_s \cdot d(p_s) \right)$$

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 $\boldsymbol{\tau}_0$-frontier.

Proposition 1. *The $\boldsymbol{\tau}_0$-frontier is the line segment of slope -1 that connects the reliance point R to the distrust point D.*

Proposition 1 shows that the efficient frontier is a line segment with constant slope -1. In this environment, neither information design nor policy design can alter the rate at which best-case and worst-case payoffs trade off. Any increase in the best-case payoff necessarily comes at an equal reduction in the worst-case payoff. While the exact slope depends on the relative costs of false negatives and false positives, the key point is that it is constant: there is no curvature to exploit, and no design choice can relax this fundamental tension.

The two endpoints of the frontier correspond to extreme implementation choices. At the distrust point D, the designer effectively shuts the AI out by collapsing the auxiliary covariate X to a single value, eliminating both upside and downside from delegation. At the reliance point R, the designer fully exposes herself to the AI by allowing covariates that, in the best case, enable perfect targeting within each subgroup, but that also permit maximal harm under adversarial behavior.

Because the frontier is linear, the optimal choice depends only on the relative weight placed on worst- versus best-case outcomes. Designers with $\eta > 1/2$ optimally select the distrust point, while designers with $\eta < 1/2$ optimally select the reliance point. No intermediate choice strictly improves upon these extremes.

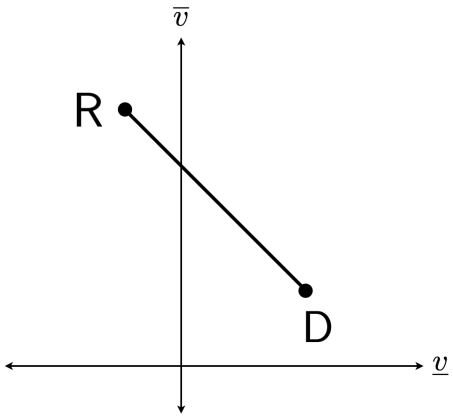


Figure 3: The efficient frontier is a line segment with slope -1.

5 General Case

We now characterize the efficient frontier in our full model. Section 5.1 generalizes the benchmark payoffs from Section 4.1. Section 5.2 characterizes the $\boldsymbol{\tau}$-frontier for a fixed but arbitrary $\boldsymbol{\tau}$, thus generalizing Proposition 1. Section 5.3 then uses this result to solve for the optimal design of $\boldsymbol{\tau}$. Section 5.4 uses the above to characterize the resulting frontier.

5.1 Generalization of Benchmark Payoffs

Let π denote the joint distribution of the binary random variables (A, Y). Define

$$\Pi_\tau := \{\pi \in \Delta(\{0,1\} \times \{0,1\}) : \underline{\tau} \leq \pi(Y = 1 \mid A = a) \leq \overline{\tau} \quad \forall a \in \{0,1\}\}$$

to be the set of all joint distributions for (A, Y) that satisfy the τ constraints.

Definition 9 (Constrained Best and Worst Targeting). Fix $p, q \in [0,1]$ and $\tau = (\underline{\tau}, \overline{\tau}) \in [0,1] \times [0,1]$. Let π range over Π_τ with marginals $\pi(A = 1) = q$ and $\pi(Y = 1) = p$. The *τ-constrained best-targeting payoff* is

$$b_p(q; \tau) := \sup_{\pi \in \Pi_\tau} \mathbb{E}_\pi\big[u(A, Y)\big],$$

and the *τ-constrained worst-targeting payoff* is

$$w_p(q;\tau) := \inf_{\pi \in \Pi_\tau} \mathbb{E}_\pi\big[u(A,Y)\big].$$

This definition is the analog of Definition 5 where the conditional success rate is constrained to lie in $[\underline{\tau}, \overline{\tau}]$. When $\tau = (0,1)$, then $b_p(q;\tau)$ and $w_p(q;\tau)$ reduce to the original $b_p(q)$ and $w_p(q)$.

5.2 General $\boldsymbol{\tau}$-Efficient Frontier

Fix a vector of informativeness bounds $\boldsymbol{\tau} = (\tau_s)_{s\in\mathcal{S}}$. We first characterize the efficient frontier for a given subgroup s, and then aggregate across subgroups.

Within subgroup s, the designer knows that a fraction p_s of patients benefit from treatment, and (given $\boldsymbol{\tau}$) has posterior beliefs constrained to within $[\underline{\tau}_s, \overline{\tau}_s]$. These informativeness bounds restrict how much the covariates can concentrate probability mass at extreme posteriors. In particular,

$$q_s := \frac{p_s - \underline{\tau}_s}{\overline{\tau}_s - \underline{\tau}_s},$$

is the largest share of subgroup s that can be assigned posterior belief $\overline{\tau}_s$ while remaining consistent with the baseline probability p_s. This quantity will play a central role below, and we use

$$w_s(\boldsymbol{\tau}) := w_{p_s}\big(q_s;\tau_s\big), \qquad b_s(\boldsymbol{\tau}) := b_{p_s}\big(q_s;\tau_s\big).$$

to denote the corresponding worst- and best-case targeting payoffs induced by the bounds τ_s.

Definition 10. The *subgroup s reliance point* induced by bounds τ_s is

$$R_s(\boldsymbol{\tau}) := \big(w_s(\boldsymbol{\tau}),\, b_s(\boldsymbol{\tau})\big),$$

and the *subgroup s distrust point* is

$$D_s := (d_s, d_s).$$

Like the trust and distrust points defined in the previous section, these points will define the endpoints of the efficient frontier.

Lemma 2. *Fix any subgroup s. Then:*

(a) If $\underline{\tau}_s > 1/2$ or $\overline{\tau}_s < \frac{1}{2}$, the subgroup s frontier is simply the distrust point D_s.

(b) Otherwise, the efficient frontier for subgroup s is the line segment that connects the subgroup s reliance point $R_s(\boldsymbol{\tau})$ to the distrust point D_s.

If $\overline{\tau} < 1/2$ then the AI cannot induce the designer to assign probability greater than $1/2$ to need of treatment, so the designer never treats. And if $\underline{\tau} < 1/2$, then the AI cannot induce the designer to assign probability less than $1/2$ to need of treatment, so the designer always treats. In either case, the subgroup frontier collapses to the distrust point D_s (part (a)). Otherwise, the subgroup frontier is the line segment connecting $R_s(\boldsymbol{\tau})$ and D_s (part (b)).

We aggregate these subgroup frontiers into an overall efficient frontier by taking the Minkowski sum of the subgroup frontiers, weighted by the subgroup distribution μ. The aggregate frontier takes the following form. Let $\Delta(s)$ denote the slope of the line connecting $R_s(\boldsymbol{\tau})$ to D_s, which can be interpreted as the tradeoff between the worst-case and best-case payoff. Order the elements of $\mathcal{S}$ as $s^{(1)}, \ldots, s^{(k)}$ where

$$\Delta(s^{(1)}) \leq \cdots \leq \Delta(s^{(k)})$$

breaking ties arbitrarily. In this ordering, the segment connecting $R_{s^{(j)}}(\boldsymbol{\tau})$ to $D_{s^{(j)}}$ is steepest at $j = 1$ and becomes progressively flatter as j increases; equivalently, the marginal cost of improving the best-case outcome in subgroup $s^{(j)}$ is increasing in j. For convenience, write $\Delta_j := \Delta(s^{(j)})$ and let $S_j = \left\{s^{(1)}, \ldots, s^{(j)}\right\}$ denote the first j subgroups in this ordering.

We construct extreme frontier points by combining trust and distrust across subgroups. For each $j = 1, \ldots, k$, define the j-th reliance point to be

$$\mathsf{R}^{(j)} = \sum_{s \in S_j} \mu_s \cdot R_s(\boldsymbol{\tau}) + \sum_{s \notin S_j} \mu_s \cdot D_s.$$

To obtain $\mathsf{R}^{(j)}$, the designer implements the reliance point $R_s(\boldsymbol{\tau})$ for each subgroup $s \in S_j$, and the distrust point D_s for every other subgroup. Intuitively, $\mathsf{R}^{(j)}$ is the outcome obtained by relying on the AI in the j subgroups where doing so has the smallest worst-case cost, and reverting to the baseline in the remaining subgroups.

Definition 11. For any two points $A, B \in \mathbb{R}^2$ let $\overline{AB}$ denote the line segment connecting these points, i.e., $\overline{AB} = \{tA + (1-t)B : t \in [0,1]\}$.

Theorem 1. *The $\boldsymbol{\tau}$-frontier is*

$$\overline{\mathsf{D}\mathsf{R}^{(1)}} \cup \overline{\mathsf{R}^{(1)}\mathsf{R}^{(2)}} \cup \overline{\mathsf{R}^{(2)}\mathsf{R}^{(3)}} \cup \cdots \cup \overline{\mathsf{R}^{(k-1)}\mathsf{R}^{(k)}}$$

This result generalizes Proposition 1 and coincides with it when $\boldsymbol{\tau} = \boldsymbol{\tau}_0$. In that case, every segment of the frontier has the same slope -1, so the frontier can be written as the single line segment $\overline{\mathsf{D}\mathsf{R}^{(k)}}$. Moreover $\mathsf{R}^{(k)}$ is precisely the reliance point R defined in Section 4.

5.3 Optimal choice of τ

We now ask how a designer with preferences $\eta\,\underline{v} + (1-\eta)\,\overline{v}$ optimally chooses the bounds $\boldsymbol{\tau}$. In each subgroup, the optimal bounds exhibit a cutoff structure in η: for extreme preferences the designer chooses corner solutions, while for intermediate preferences the bounds adjust smoothly.

Proposition 2. *Fix $\eta \in [0,1]$. For each subgroup s, there are thresholds $0 < \underline{\eta}_s < \overline{\eta}_s < 1$ such that an optimal choice of $(\underline{\tau}_s, \overline{\tau}_s)$ has the following form:*

(a) If $p_s \leq \frac{1}{2}$, then $\overline{\tau}_s^* = 1$ for all η, and $\underline{\tau}_s^*$ equals 0 for $\eta \leq \underline{\eta}_s$, equals p_s for $\eta \geq \overline{\eta}_s$, and increases monotonically from 0 to p_s as η ranges from $\underline{\eta}_s$ to $\overline{\eta}_s$.

(b) If $p_s > \frac{1}{2}$, then $\underline{\tau}_s^* = 0$ for all η, and $\overline{\tau}_s^*$ equals 1 for $\eta \leq \underline{\eta}_s$, equals p_s for $\eta \geq \overline{\eta}_s$, and decreases monotonically from 1 to p_s as η ranges from $\underline{\eta}_s$ to $\overline{\eta}_s$.

This proposition emphasizes the qualitative nature of the frontier; a complete characterization can be found in Proposition B.1 in the appendix. Note that by Bayes

plausibility, setting $(\underline{\tau}_s, \overline{\tau}_s) = (p_s, 1)$ or $(\underline{\tau}_s, \overline{\tau}_s) = (0, p_s)$ is equivalent to setting $(\underline{\tau}_s, \overline{\tau}_s) = (p_s, p_s)$, because the posterior is constrained to equal to the prior.

For intuition, consider the case $p_s \leq \frac{1}{2}$; the case $p_s > \frac{1}{2}$ is symmetric. For any choice of bounds τ_s, the subgroup-s frontier is a line segment connecting the distrust point $D_s = (0, 0)$ to the reliance point $R_s(\boldsymbol{\tau}) = (w_s(\boldsymbol{\tau}), b_s(\boldsymbol{\tau}))$. Thus the designer's choice of τ_s affects outcomes only through the location of $R_s(\boldsymbol{\tau})$ and the slope of the segment connecting D_s to $R_s(\boldsymbol{\tau})$.

The designer would like the reliance point to be as far to the "north" as possible. This consideration pushes toward permissive bounds: lowering $\underline{\tau}_s$ and raising $\overline{\tau}_s$ allow an aligned AI to concentrate treatment on patients most likely to benefit, thereby improving the best-case outcome.

At the same time, the designer would like for the reliance point to be as far "east" as possible, corresponding to a better tradeoff between worst- and best-case payoffs: sacrificing a small amount of worst-case performance yields a relatively large improvement in the best case. Both bounds matter for this slope. Raising $\overline{\tau}_s$ improves the best case by allowing more precise targeting, while raising $\underline{\tau}_s$ improves the worst case by limiting how poorly a misaligned AI can perform—even in the worst case, at least a fraction $\underline{\tau}_s$ of treated patients truly benefit.

These considerations generate an asymmetry. Increasing $\overline{\tau}_s$ is unambiguously beneficial: it improves the best-case payoff and steepens the frontier. In contrast, increasing $\underline{\tau}_s$ involves a genuine tradeoff. Raising $\underline{\tau}_s$ limits the scope for a misaligned AI and thereby improves worst-case performance, but it also constrains an aligned AI's ability to target effectively, reducing the best-case payoff.

The optimal lower bound $\underline{\tau}_s^*$ therefore balances these forces. 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, $\tau_s = (0, 1)$, or imposing full restrictions $\tau_s = (p_s, p_s)$. For intermediate values of η, the optimal $\underline{\tau}_s^*$ is interior and adjusts smoothly with preferences. When $p_s > \frac{1}{2}$, the logic is symmetric but reversed: the key role of the bounds is no longer to identify whom to treat, but rather whom to avoid.

5.4 The Endogenous Frontier

We now characterize the efficient frontier under the designer's optimal choice of informativeness bounds $\boldsymbol{\tau}$. The resulting frontier has a simple cutoff structure, where subgroups are ranked by how balanced they are, $|p_s - \frac{1}{2}|$. For any preference parameter η, the designer relies on the AI for the initial subgroups in this ordering and distrusts it for the remaining subgroups, with the location of this cutoff varying monotonically with η.

Definition 12. Order the elements of $\mathcal{S}$ as $s^{(1)}, \ldots, s^{|\mathcal{S}|}$ such that

$$i < j \quad \Longleftrightarrow \quad \left|p_{s^{(i)}} - \tfrac{1}{2}\right| \leq \left|p_{s^{(j)}} - \tfrac{1}{2}\right|.$$

Proposition 3 (Cutoff structure of optimal trust)**.** *For every $\eta \in [0,1]$, there exists a cutoff index $J_\eta \in \{0, \ldots, |\mathcal{S}|\}$ such that the designer optimally implements the trust point for subgroup $s^{(j)}$ if and only if $j \leq J_\eta$. Moreover, J_η is weakly increasing in η.*

To understand this cutoff, consider how the designer's optimal choice of $\boldsymbol{\tau}$ changes as the weight on worst-case performance varies. When $\eta > \max_{s \in \mathcal{S}} \overline{\eta}_s$, the designer places overwhelming weight on worst-case payoff and thus optimally chooses a completely constrained information environment in which the AI has no scope to influence the designer's decisions. This results in a frontier that is simply the single distrust point D.

As η decreases, the designer begins to allow informativeness selectively. The first subgroup for which bounds are relaxed is the one with the smallest $|p_s - \frac{1}{2}|$. In this subgroup, a misaligned AI has the least ability to cause harm, so the designer faces a better tradeoff between the best-case and worst-case payoffs. Relaxing τ_s for this subgroup creates a nondegenerate subgroup frontier, and the designer optimally selects its reliance point. 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 subgroups enter the reliance set, so their frontiers also become nondegenerate. Second, for subgroups where AI is already being relied on, the bounds are relaxed further.

Both adjustments improve the best-case payoff by allowing more precise targeting, but at the cost of a flatter worst–best tradeoff. The optimal balance between these effects determines how many subgroups are trusted at any given η.

Finally, when $\eta < \min_{s\in\mathcal{S}} \underline{\eta}_s$, the designer optimally chooses fully permissive bounds $\tau_s = (0, 1)$ for every subgroup. In this limit, the efficient frontier coincides with the benchmark frontier characterized in Section 4.

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 subgroups by age, with the following table denoting the baseline probability that the treatment is needed in each subgroup. 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 3.

Table 1: Optimal trust by subgroup

s	p_s	$\lvert p_s - \frac{1}{2}\rvert$	$\eta < \eta^{(1)}$	$\eta^{(1)} \le \eta < \eta^{(2)}$	$\eta^{(2)} \le \eta < \eta^{(3)}$	$\eta \ge \eta^{(3)}$
18–39	0.52	0.02	**Trust**	**Trust**	**Trust**	**Distrust**
40–75	0.70	0.20	**Trust**	**Trust**	**Distrust**	**Distrust**
75+	0.10	0.40	**Trust**	**Distrust**	**Distrust**	**Distrust**

Figure 4 illustrates how the efficient 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 subgroup. As η increases past $\eta^{(1)}$, the hospital first withdraws trust in the 75+ subgroup (the group with largest $|p_s - 0.5|$), 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 subgroup, so that the AI is trusted only for the 18-39 group. Finally, when $\eta \ge \eta^{(3)}$, worst case concerns are strong enough that the hospital distrusts the AI for all subgroups.

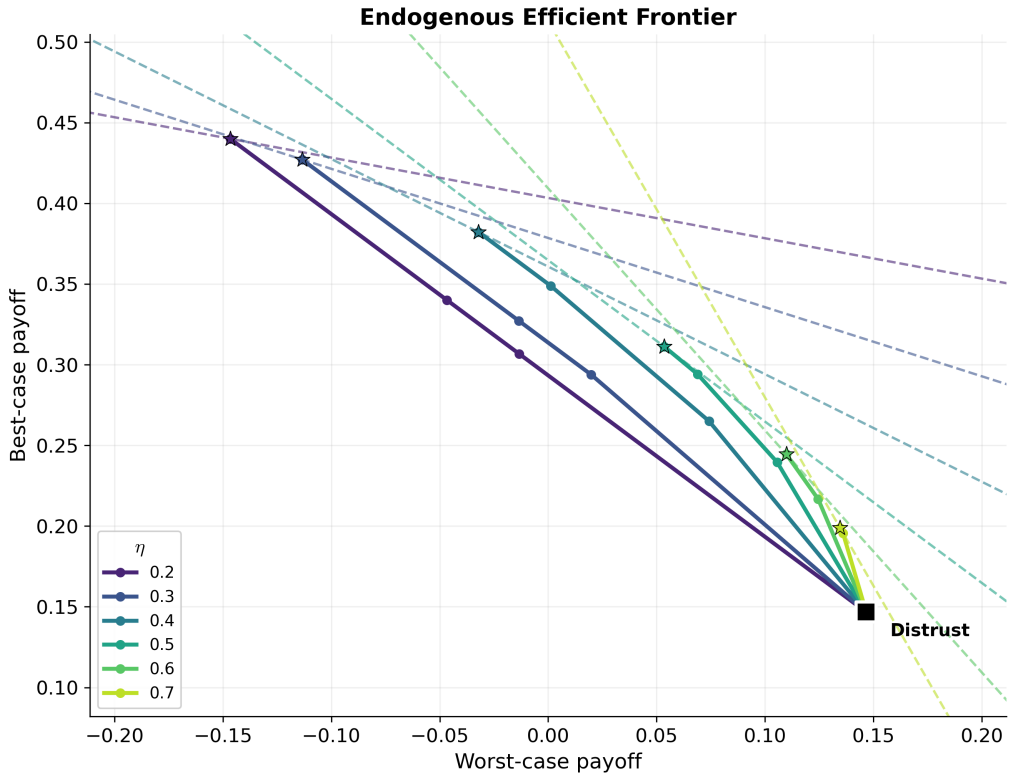


Figure 4: The solid curves depict the frontier (with optimal informativeness bounds) for different η's. bounds $(\underline{\tau}_s, \overline{\tau}_s)$. The frontiers "fan out" from the distrust point as η decreases: lower η (more weight on best-case payoff) yields frontiers that extend further toward high $\overline{v}$ but with steeper slopes. Stars indicate the optimal point on each frontier, and dashed lines show the supporting hyperplanes.

6 Conclusion

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" and explore what its consequences would be. Second, we conduct our analysis in the simplest possible decision setting: a one-time decision of which patients to treat. Future work could consider repeated interaction with an AI and more general state spaces and payoff functions. Third, we suppose that Nature's choice of distribution (i.e., the true relationship between attributes and treatment need) is as favorable or unfavorable as possible. Future work could relax

this assumption by constraining the space of conceivable distributions given other observables, for example by supposing that our informativeness bounds $\underline{\tau}$ and $\overline{\tau}$ are a function of the richness of the covariate space $\mathcal{X}$.

A Proof of Theorem 1

A.1 Preliminary Results

Lemma A.1. *Fix any $p, q \in [0,1]$ and $\tau = (\underline{\tau}, \overline{\tau}) \in [0,1] \times [0,1]$. Let*

$$q^* = \frac{p - \underline{\tau}}{\overline{\tau} - \underline{\tau}}.$$

Then

$$b_p(q;\tau) = \begin{cases} q(2\overline{\tau} - 1) & \text{if } q \leq q^* \\ 2p - 2(1-q)\underline{\tau} - q & \text{if } q > q^* \end{cases}$$

$$w_p(q;\tau) = \begin{cases} q(2\underline{\tau} - 1) & \text{if } q \leq 1 - q^* \\ 2p - 2(1-q)\overline{\tau} - q & \text{if } q > 1 - q^*. \end{cases}$$

Corollary A.1. *Fix any subgroup s and let*

$$q_s := \frac{p_s - \underline{\tau}_s}{\overline{\tau}_s - \underline{\tau}_s}$$

Then

$$b(s) := b_{p_s}(q_s, \tau_s) = q_s \cdot (2\overline{\tau}_s - 1)$$

$$w(s) := w_{p_s}(q_s, \tau_s) = \begin{cases} q_s \cdot (2\underline{\tau}_s - 1) & \text{if } q_s \leq 1/2 \\ 2(p_s - \overline{\tau}) + q_s \cdot (2\overline{\tau} - 1) & \text{if } q_s > 1/2. \end{cases}$$

Proof. 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 $\pi(A = 1) = q$ is fixed, the objective reduces to $q(2\pi_1 - 1)$, 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 for $p = \pi(Y = 1)$ implies

$$q\pi_1 + (1 - q)\pi_0 = p. \tag{A.1}$$

Second, τ-admissibility of π requires

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

If $q = 0$, the designer's payoff is 0, which is consistent with the expressions in the result. If $q > 0$ then (from (A.1))

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

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

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

The feasible set for π_1 is then the closed interval

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

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

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

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

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

Thus

$$b_p(q;\tau) = \begin{cases} q(2\overline{\tau}-1), & q \le q^*, \\ 2p - 2(1-q)\underline{\tau} - q, & q > q^*. \end{cases}$$

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

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

Therefore,

$$w_p(q;\tau) = \begin{cases} q(2\underline{\tau}-1), & q \le 1-q^*, \\ 2p - 2(1-q)\overline{\tau} - q, & q > 1-q^* \end{cases}.$$

□

A.2 Subgroup Frontier

Fix a choice of (I,σ) and a subgroup $s \in \mathcal{S}$. Let $p := p_s$ denote the (known) fraction of patients in the subgroup who need treatment, and let ν denote the conditional distribution of $X \mid S = s$. All random variables below are defined on the probability space $(\mathcal{X}, \mathcal{F}, \nu)$.

Since $u(0,Y) = 0$ and $u(1,Y) = 2Y - 1$, the designer's payoff can be written as

$$\mathbb{E}[A(2Y-1)],$$

where Y is the conditional probability of treatment need and A is the probability of treatment. The $\boldsymbol{\tau}$-admissibility constraints imply that Y may be any random variable satisfying

$$\underline{\tau}_s \le Y \le \overline{\tau}_s \qquad \text{and} \qquad \mathbb{E}[Y] = p.$$

Let $\mathbb{Y}$ denote the collection of all such random variables. The designer's strategy σ restricts the set of implementable treatment rules to a collection $\mathbb{A}_\sigma$ of random variables taking values in $[0,1]$. (Specifically, $A \in \mathbb{A}_\sigma$ if and only if there exists a report $P \in \mathcal{P}$ such that $A(x) = \sigma(P)(s,x)$ for every $x \in \mathcal{X}$.)

We can then define the worst- and best-case subgroup payoffs as

$$\underline{v} = \inf_{Y \in \mathbb{Y}} \inf_{A \in \mathbb{A}_\sigma} \mathbb{E}_\nu[A(2Y-1)],$$
$$\overline{v} = \sup_{Y \in \mathbb{Y}} \sup_{A \in \mathbb{A}_\sigma} \mathbb{E}_\nu[A(2Y-1)].$$

Thus, within each subgroup, the problem reduces to an extremal problem over pairs of random variables (A, Y) with fixed marginal constraints.

We derive three bounds on the feasible set of worst- and best-case payoff pairs $(\underline{v}, \overline{v})$, showing they lie below the line segment $\overline{\mathsf{RD}}$ as depicted in Figure 5. We then construct explicit information environments and policies that attain every point on this segment, proving these bounds are tight. Finally, we aggregate the subgroup-level frontiers by taking their weighted Minkowski sum.

A.2.1 Bounds on the Feasible Set

Since s and $\boldsymbol{\tau}$ are fixed, write

$$b := b_s(\boldsymbol{\tau}), \qquad d := d_s, \qquad w := w_s(\boldsymbol{\tau}).$$

Lemma A.2. *The best case payoff satisfies $\overline{v}_s \leq b$.*

Lemma A.3. *The worst case payoff satisfies $\underline{v}_s \leq d$.*

Lemma A.4. *Let $c_1 := b - d$, $c_2 := d - w$, and $c_3 := (b - w)d$. Then every feasible payoff pair satisfies*

$$c_1 \cdot \underline{v} + c_2 \cdot \overline{v} \leq c_3.$$

Taken together, Lemmas A.2-A.4 confine 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 $\overline{\mathsf{RD}}$. Crucially, these bounds depend only on (p_s, τ_s) and not on the particular choice of (ν, σ). See the shaded region of Figure 5.

The *trust point*

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

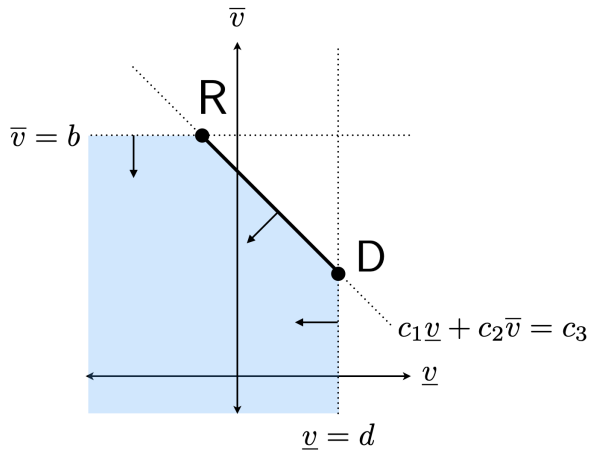


Figure 5: Any feasible $(\underline{v}, \overline{v})$ falls in the shaded region.

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 this line segment is implementable, in which case it will be the efficient frontier. To implement D, choose any X and let A be the constant random variable $A = \mathbb{1}(p \geq 1/2)$. Then $\mathbb{E}[A(2Y-1)] = (2p-1)_+ = d$. To implement R, choose a binary covariate with $\Pr(X = 1) = q^* := \frac{p-\tau}{\overline{\tau}-\underline{\tau}}$. Choose the decision rule that "follows the AI's report" but caps treatment at a fraction q^* of patients: That is, if $\nu\big(\{x : \ P(Y = 1 \mid S = s, X = x) \geq \frac{1}{2}\}\big) \leq q^*$ then

$$\alpha_P(s, x) = \mathbb{1}(P(Y = 1 \mid S = s, X = x) \geq 1/2)$$

and otherwise let $\alpha_P(s, \cdot)$ treat exactly a ν-mass q^* of patients, chosen from among those x with the largest reported values of $P(Y = 1 \mid S = s, X = x)$ (ties broken

arbitrarily).[5]

We claim that under this policy, the set of joint distributions over (A, Y) induced by I-admissible distributions coincides with Π_τ with marginals $\pi(A = 1) = q^*$ and $\pi(Y = 1) = p$. Indeed, fix any admissible distribution P^* and any AI report P. Conditional on $S = s$, the action $A = \alpha_P(s, X)$ is a $[0, 1]$-valued function of X, hence it induces some treatment rate $q := \mathbb{E}_\nu[A] \leq q^*$ and a joint law π over (A, Y) satisfying

$$\underline{\tau} \leq \pi(Y = 1 \mid A = a) \leq \overline{\tau} \qquad \forall a \in \{0, 1\}, \qquad \pi(Y = 1) = p.$$

Conversely, because X is chosen by the designer and $\nu(X = 1) = q^*$, for any $\pi \in \Pi_\tau$ with $\pi(A = 1) = q^*$ and $\pi(Y = 1) = p$, we can realize π by taking $X \in \{0, 1\}$ with $\nu(X = 1) = q^*$ and setting

$$P^*(Y = 1 \mid S = s, X = 1) = \pi(Y = 1 \mid A = 1)$$
$$P^*(Y = 1 \mid S = s, X = 0) = \pi(Y = 1 \mid A = 0)$$

and choosing a report P for which the policy selects $A = \mathbb{1}(X = 1)$. (For example, the AI can report posteriors above $1/2$ on $X = 1$ and below $1/2$ on $X = 0$, which is feasible whenever $\underline{\tau} \leq \frac{1}{2} \leq \overline{\tau}$.)

Therefore, the best- and worst-case payoffs induced by this construction are exactly the extrema of $\mathbb{E}_\pi[u(A, Y)]$ over Π_τ with marginals $\pi(A = 1) = q^*$ and $\pi(Y = 1) = p$. By Lemma A.1, these extrema are

$$\overline{v}_{\mathsf{R}} = b_p(q^*; \tau) = b, \qquad \underline{v}_{\mathsf{R}} = w_p(q^*; \tau) = w,$$

so the resulting payoff pair is precisely the reliance point $\mathsf{R} = (w, b)$.

Finally, since ex ante randomization over feasible (I, σ) convexifies the set of achievable payoff pairs, every point on the line segment $\overline{\mathsf{RD}}$ is implementable by mixing between the constructions that implement R and D. Hence $\overline{\mathsf{RD}}$ is the (subgroup) efficient frontier.

[5]Equivalently, $\alpha_P(s, \cdot)$ is a measurable selector that maximizes $\mathbb{E}_\nu[\alpha_P(s, X) \cdot P(Y = 1 \mid S = s, X)]$ subject to $\mathbb{E}_\nu[\alpha_P(s, X)] \leq q^*$.

A.2.2 Proof of Lemma A.2

For each treatment probability $q \in [0,1]$, let

$$\mathbb{A}_q = \{A \in \mathbb{A}_\sigma : \mathbb{E}[A] = q\}$$

denote those implementable action variables with a treatment rate of q. Then

$$\overline{v} = \sup_{q\in[0,1]} \sup_{Y\in\mathbb{Y}} \sup_{A\in\mathbb{A}_q} \mathbb{E}[A(2Y-1)]$$

But conditional on this subgroup, $\sup_{Y\in\mathbb{Y}} \sup_{A\in\mathbb{A}_q} \mathbb{E}[A(2Y-1)]$ is the same extremal problem over (A,Y) as in the definition of $b_p(q;\tau)$, up to the restriction that A be implementable. Thus we have

$$\overline{v} \leq \sup_{q\in[0,1]} b_p(q;\tau).$$

For any fixed (p,τ), the function $b_p(q;\tau)$ is piecewise linear in q with a single kink at $q = q^*$ (Lemma A.1). Moreover, it is increasing on $[0,q^*]$ and weakly decreasing on $[q^*,1]$. Therefore,

$$\sup_{q\in[0,1]} b_p(q;\tau) = b_p(q^*;\tau) = b$$

proving the result.

A.2.3 Proof of Lemma A.3

Let $Y \equiv p$ almost surely. Since $p \in [\underline{\tau}_s, \overline{\tau}_s]$, this choice is $\boldsymbol{\tau}$-admissible. Then for any implementable action variable A,

$$\begin{aligned}
\mathbb{E}[A(2Y-1)] &= \mathbb{E}[A]\,(2p-1) \\
&\leq (2p-1)_+ && \text{since } A \in [0,1] \\
&= d
\end{aligned}$$

Since Nature can choose this Y, it follows that the worst-case payoff must satisfy $\underline{v} \leq d$.

A.2.4 Proof of Lemma A.4

We prove the proposition by first establishing a pointwise bound for each implementable action variable. For any $A \in [0,1]$, define

$$\underline{u}(A) := \inf_{Y \in \mathbb{Y}} \mathbb{E}[A(2Y-1)],$$
$$\overline{u}(A) := \sup_{Y \in \mathbb{Y}} \mathbb{E}[A(2Y-1)]$$

to be the worst- and best-case payoffs against Nature for this fixed choice of A. We will show that

$$c_1 \cdot \underline{u}(A) + c_2 \cdot \overline{u}(A) \leq c_3 \tag{A.3}$$

with c_1, c_2, c_3 as defined in the statement of the proposition, and then argue that this implies the desired result.

Write $Y = \underline{\tau} + (\overline{\tau} - \underline{\tau})Z$, where Z is a random variable satisfying $0 \leq Z \leq 1$. The constraint $\mathbb{E}[Y] = p$ is equivalent to

$$\mathbb{E}[Z] = \frac{p - \underline{\tau}}{\overline{\tau} - \underline{\tau}} =: q^*.$$

Let $\mathbb{Z}$ be the set of random variables Z satisfying this range and mean constraint. Then

$$\begin{aligned}
\underline{u}(A) &= \inf_{Y \in \mathbb{Y}} \mathbb{E}[A(2Y-1)] \\
&= \inf_{Z \in \mathbb{Z}} \mathbb{E}\big[A\big(2(\underline{\tau} + (\overline{\tau} - \underline{\tau})Z) - 1\big)\big] \\
&= (2\underline{\tau} - 1)\mathbb{E}[A] + 2(\overline{\tau} - \underline{\tau}) \inf_{Z \in \mathbb{Z}} \mathbb{E}[AZ].
\end{aligned}$$

Since the objective is linear in Z and the feasible set $\{Z : \ 0 \leq Z \leq 1, \ \mathbb{E}[Z] = q^*\}$ is convex, the infimum is attained at an extreme point, i.e. by a $\{0,1\}$-valued Z with $\mathbb{E}[Z] = q^*$. Among such Z, the minimum of $\mathbb{E}[AZ]$ is achieved by taking $Z = 1$ on the lowest q^*-quantile of A (a maximally countermonotone coupling), yielding

$$\inf_{Z \in \mathbb{Z}} \mathbb{E}[AZ] = \int_0^{q^*} Q_A(u)\, du,$$

where Q_A is a left-continuous quantile function of A. Therefore

$$\underline{u}(A) = (2\underline{\tau} - 1)\mathbb{E}[A] + 2(\overline{\tau} - \underline{\tau}) \int_0^{q^*} Q_A(u)\, du.$$

By identical arguments, the best case payoff is

$$\overline{u}(A) = (2\underline{\tau} - 1)\mathbb{E}[A] + 2(\overline{\tau} - \underline{\tau}) \int_{1-q^*}^{1} Q_A(u) du$$

Thus for $c_1 = b - d$ and $c_2 = d - w$,

$$\Lambda(A) := c_1 \underline{u}(A) + c_2 \overline{u}(A)$$

is a linear functional of the quantile function Q_A. The set of distributions of $[0,1]$-valued random variables with a fixed mean is convex, and its extreme points are distributions supported on $\{0,1\}$. Therefore, it suffices to consider $\{0,1\}$-valued treatment rules A_q that treat exactly a q-fraction of the population. That is,

$$\sup_{A \in \mathbb{A}_\sigma} \Lambda(A) \leq \sup_{q \in [0,1]} \Lambda(A_q)$$

where A_q denotes the action variable that treats a q fraction of the population.

Thus we can restrict our attention to upper bounding

$$\Lambda(A_q) = (b - w)(2\underline{\tau} - 1)q + 2(\overline{\tau} - \underline{\tau})\left((b - d)T(q) + (d - w)B(q)\right)$$

where

$$T(q) = \int_0^{q^*} Q_{A_q}(u) du = \begin{cases} 0 & \text{if } q^* \leq 1 - q \\ q^* - (1 - q) & \text{if } q^* > 1 - q \end{cases}$$

$$B(q) = \int_{1-q^*}^{1} Q_{A_q}(u) du = \begin{cases} q^* & \text{if } q^* \leq q \\ q & \text{if } q^* > q \end{cases}$$

The function $\Lambda(A_q)$ is piecewise linear in q, so it is maximized at one of the kink points $q \in \{0, q^*, 1 - q^*, 1\}$. Plugging these values into the expressions above, a direct

comparison shows that the maximizer is $q = q^*$, given which

$$(\underline{u}(A_{q^*}), \overline{u}(A_{q^*})) = (w, b) \quad \Rightarrow \quad \Lambda(A_{q^*}) = (b-d)w + (d-w)b = (b-w)d.$$

Thus

$$(b-d)\underline{u}(A) + (d-w)\overline{u}(A) \leq \sup_{A \in \mathbb{A}_\sigma} \Lambda(A) \leq \sup_{q \in [0,1]} \Lambda(A_q) = (b-w)d$$

as desired.

It remains to show that the pointwise bound $c_1\underline{u}(A) + c_2\overline{u}(A) \leq c_3$ for every $A \in \mathbb{A}_\sigma$ implies the stated bound for $(\underline{v}, \overline{v})$. Recall that

$$\underline{v} = \inf_{A \in \mathbb{A}_\sigma} \underline{u}(A), \qquad \overline{v} = \sup_{A \in \mathbb{A}_\sigma} \overline{u}(A).$$

Fix $\varepsilon > 0$ and choose $A_\varepsilon \in \mathbb{A}_\sigma$ such that

$$\overline{u}(A_\varepsilon) \geq \overline{v} - \varepsilon.$$

Then also $\underline{v} \leq \underline{u}(A_\varepsilon)$, and hence

$$c_1\underline{v} + c_2\overline{v} - c_2\varepsilon \leq c_1\underline{u}(A_\varepsilon) + c_2\overline{u}(A_\varepsilon) \leq c_3.$$

Letting $\varepsilon \to 0$ yields $c_1\underline{v} + c_2\overline{v} \leq c_3$.

A.3 Constructing the Full Frontier

We return to the full model with an arbitrary finite set $\mathcal{S}$. For each subgroup $s \in \mathcal{S}$, let $\mathsf{R}_s = (w_s(\boldsymbol{\tau}), b_s(\boldsymbol{\tau}))$ and $\mathsf{D}_s = (d_s, d_s)$ denote the subgroup trust and distrust points, and recall from the previous subsection that the subgroup efficient frontier is the line segment $\overline{\mathsf{R}_s\mathsf{D}_s}$.

Fix any feasible (I, σ). Since payoffs are expectations and S is finite with fixed marginal μ, the expected payoffs decompose as

$$\underline{v}_I(\sigma) = \sum_{s \in \mathcal{S}} \mu(s)\, \underline{v}_s(\sigma), \qquad \overline{v}_I(\sigma) = \sum_{s \in \mathcal{S}} \mu(s)\, \overline{v}_s(\sigma).$$

The closure of the global feasible set is thus the weighted Minkowski sum of the subgroup feasible sets. Formally, let $C_s := \overline{\mathsf{R}_s\mathsf{D}_s}$, and define

$$C = \sum_{s\in\mathcal{S}} \mu(s)\, C_s := \Big\{ \sum_{s\in\mathcal{S}} \mu(s)\, z_s \;\; : \;\; z_s \in C_s \Big\}. \tag{A.4}$$

The global efficient frontier is the set of undominated points in C.

Fix any payoff parameter $\eta \in [0,1]$ and consider the supporting functional

$$\Lambda_\eta(\underline{v}, \overline{v}) := \eta\, \underline{v} + (1-\eta)\, \overline{v}.$$

Because C is a weighted Minkowski sum (A.4) and Λ_η is linear, maximization separates across subgroups:

$$\max_{z\in C(I)} \Lambda_\eta(z) = \sum_{s\in\mathcal{S}} \mu(s) \max_{z_s\in C_s} \Lambda_\eta(z_s).$$

Since each C_s is a line segment, Λ_η attains its maximum on C_s at an endpoint, i.e. either at R_s or at D_s (with ties yielding the entire segment). The reliance point R_s is (weakly) preferred to the distrust D_s under Λ_η if and only if

$$\eta\, w_s(\boldsymbol{\tau}) + (1-\eta)\, b_s(\boldsymbol{\tau}) \;\geq\; d_s,$$

which is equivalent (when $w_s(\boldsymbol{\tau}) \neq d_s$) to

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

Thus, as η increases, the maximizing choice switches subgroups from R_s to D_s in weakly decreasing order of $\Delta(s)$. Therefore every maximizer of Λ_η over $C(I)$ is of the form $\mathsf{R}^{(j)}$ for some j (with ties corresponding to convex combinations between $\mathsf{R}^{(j)}$ and $\mathsf{R}^{(j+1)}$). As η varies, the set of maximizers traces exactly the chain of segments $\overline{\mathsf{R}^{(0)}\mathsf{R}^{(1)}} \cup \cdots \cup \overline{\mathsf{R}^{(k-1)}\mathsf{R}^{(k)}}$.

Finally, since C is compact and convex (a Minkowski sum of compact convex sets), every undominated boundary point is supported by some Λ_η. Hence the supported chain above coincides with the efficient frontier.

B Proof of Proposition B.1

We prove the stronger claim below.

Proposition B.1. *For any preference parameter $\eta \in [0,1]$ and any subgroup s with $p := p_s$, the optimal choice of $\tau = (\underline{\tau}, \overline{\tau}) := (\underline{\tau}_s, \overline{\tau}_s)$ for this subgroup is as follows:*

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

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

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

Proof. Because τ_s is chosen independently across subgroups and the designer's objective $\eta\underline{v} + (1-\eta)\overline{v}$ is additive across subgroups (with fixed weights $\mu(s)$), it suffices to solve the one-subgroup problem for fixed $p := p_s$.

For a fixed τ, the subgroup frontier is the line segment between the distrust point

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

and the reliance point $\mathsf{T}(\tau) = (w(\tau), b(\tau))$ where

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

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

We prove the $p \leq 1/2$ case. The distrust point $\mathsf{D} = (0, 0)$ yields a payoff of zero, and the designer can implement this by choosing $\underline{\tau} = p = \overline{\tau}$. 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. Let

$$\mathcal{T} = \{(\underline{\tau}, \overline{\tau}) \in [0, 1]^2 : 0 \leq \underline{\tau} \leq p \leq \overline{\tau} \leq 1\}$$

be the set of feasible values of $(\underline{\tau}, \overline{\tau})$.

Fix $\tau = (\underline{\tau}, \overline{\tau}) \in \mathcal{T}$ and write

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

Since $\underline{\tau} \leq p \leq \frac{1}{2}$, we have $2\underline{\tau} - 1 \leq 0$.

(A) Regime $q^* \leq \frac{1}{2}$ (equivalently $p \leq (\underline{\tau} + \overline{\tau})/2$). In this case,

$$w(\tau) = q^*(2\underline{\tau} - 1), \qquad b(\tau) = q^*(2\overline{\tau} - 1).$$

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

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

hence

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

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

$$\underline{\tau}^* = \begin{cases} 0, & \eta \le \frac{1-p}{2}, \\ 1 - \sqrt{\frac{1-p}{2\eta}}, & \frac{1-p}{2} < \eta < \frac{1}{2(1-p)}, \\ p, & \eta \ge \frac{1}{2(1-p)}, \end{cases} \qquad \text{with} \qquad \overline{\tau}^* = 1. \tag{B.1}$$

The corresponding value is

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

(B) Regime $q^* > \frac{1}{2}$ (equivalently $p > \frac{\underline{\tau}+\overline{\tau}}{2}$). By assumption, $\overline{\tau} > \frac{1}{2}$, in which case we have

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

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

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

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

With $\underline{\tau} = 0$, the condition $q^* > \frac{1}{2}$ 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

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

By the AM–GM inequality, for any $\overline{\tau} > 0$,

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

with equality at $\overline{\tau} = \sqrt{p/(2\eta)}$. Accounting for the constraint $\overline{\tau} \leq 2p$ yields

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

and the associated value is

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

We finally argue that Regime A dominates Regime B. First, for $\eta \leq \frac{1}{8p}$,

$$U_R(0,1) - U^*_B = p(1-2\eta) - \Big(2p(1-\eta) - \tfrac{1}{2}\Big) = \tfrac{1}{2} - p \geq 0,$$

so the boundary Regime B value is weakly dominated by the feasible Regime A choice $(\underline{\tau}, \overline{\tau}) = (0,1)$. Next, for $\eta \geq \frac{1}{8p}$,

$$U^*_A - U^*_B = (1-2p)(1+2\eta) - 2\sqrt{2\eta}\,(\sqrt{1-p} - \sqrt{p}).$$

For $p \in [0, \frac{1}{2}]$,

$$\sqrt{1-p} - \sqrt{p} \leq 1 - 2p, \tag{B.2}$$

so

$$U^*_A - U^*_B \geq (1-2p)\Big(1 + 2\eta - 2\sqrt{2\eta}\Big) = (1-2p)\big(\sqrt{2\eta} - 1\big)^2 \geq 0.$$

Thus the best payoff in Regime (A) is weakly higher than the best payoff in Regime (B), and the overall solution is the one given in (B.1).

The argument for $p > 1/2$ is symmetric and hence omitted. □

C Proof of Proposition 3

Corollary C.1. *Fix any preference parameter η and let $\boldsymbol{\tau}$ be as given in Proposition B.1. Then the slope of the subgroup frontier line segment $\overline{\mathsf{R}_s\mathsf{D}_s}$ is*

$$\Delta_s = \frac{b_s - d_s}{w_s - d_s}$$

where there exist $0 < \underline{\eta} < \overline{\eta} < 1$ such that

(a) if $\eta < \underline{\eta}$ then $\Delta_s = -1$,

(b) if $\underline{\eta} < \eta < \overline{\eta}$ then

$$\Delta_s = \frac{1}{1 - \sqrt{\frac{1}{\eta}(1 + 2|p - \frac{1}{2}|)}}$$

(c) if $\eta > \overline{\eta}$ the line segment is degenerate.

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

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

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

since $p \leq \frac{1}{2}$ implies $1 + 2|p - \frac{1}{2}| = 2(1 - p)$. This yields Part (b).

Finally, when $\eta > \overline{\eta}$ then the designer optimally chooses $(\underline{\tau}, \overline{\tau}) = (p, p)$ so $\mathsf{R}_s = \mathsf{D}_s = (0, 0)$.

The argument for $p > 1/2$ is symmetric.

□

This corollary implies that for each subgroup s there exist thresholds $\underline{\eta}_s := \underline{\eta}(p_s)$

and $\overline{\eta}_s := \overline{\eta}(p_s)$ such that: (i) if $\eta < \underline{\eta}_s$ then $\Delta(s,\eta) = -1$; (ii) if $\underline{\eta}_s < \eta < \overline{\eta}_s$ then

$$\Delta(s,\eta) = \frac{1}{1 - \sqrt{\frac{1}{\eta}\big(1 + 2|p - \frac{1}{2}|\big)}}; \tag{C.1}$$

(iii) if $\eta > \overline{\eta}_s$ the segment is degenerate. Moreover, the thresholds themselves can be written as functions of $g_s := |p_s - \frac{1}{2}|$:

$$\underline{\eta}_s = \frac{1}{4} + \frac{g_s}{2} = \frac{1 + 2g_s}{4}, \qquad \overline{\eta}_s = \frac{1}{1 + 2g_s}.$$

In particular, $\underline{\eta}_s$ is weakly increasing in g_s and $\overline{\eta}_s$ is weakly decreasing in g_s.

Fix η and consider two subgroups s, s' with $g_s \leq g_{s'}$. We claim $\Delta(s,\eta) \leq \Delta(s',\eta)$. If either segment is degenerate, the claim holds by the convention that degenerate segments have slope $-\infty$. Otherwise, there are two cases.

Case A: $\eta \leq \underline{\eta}_s$. Then $\eta \leq \underline{\eta}_{s'}$ as well because $\underline{\eta}_s$ is increasing in g_s. Hence $\Delta(s,\eta) = \Delta(s',\eta) = -1$.

Case B: $\eta > \underline{\eta}_s$. If $\eta \geq \overline{\eta}_s$ then s is degenerate and we are done; so suppose $\underline{\eta}_s < \eta < \overline{\eta}_s$. If also $\underline{\eta}_{s'} < \eta < \overline{\eta}_{s'}$, then (C.1) applies for both subgroups. For fixed η, the map

$$\phi_\eta(g) := \frac{1}{1 - \sqrt{\frac{1}{\eta}(1 + 2g)}}.$$

is strictly increasing in g on the region where it is defined (i.e. where $\underline{\eta}(g) < \eta < \overline{\eta}(g)$). Therefore $\Delta(s,\eta) = \phi_\eta(g_s) \leq \phi_\eta(g_{s'}) = \Delta(s',\eta)$.

If instead s' is not in the interior region (i.e. $\eta \leq \underline{\eta}_{s'}$ or $\eta \geq \overline{\eta}_{s'}$), then $\Delta(s',\eta)$ equals -1 or corresponds to a degenerate segment. In either case, the inequality $\Delta(s,\eta) \leq \Delta(s',\eta)$ still holds since $\Delta(s,\eta) \leq -1$.[6] This proves the monotonicity claim:

$$g_s \leq g_{s'} \quad \Longrightarrow \quad \Delta(s,\eta) \leq \Delta(s',\eta). \tag{C.2}$$

We finally order subgroups as $s^{(1)}, \ldots, s^{(|\mathcal{S}|)}$ such that

$$g_{s^{(1)}} \leq g_{s^{(2)}} \leq \cdots \leq g_{s^{(|\mathcal{S}|)}}.$$

[6]Indeed, in the interior region $\underline{\eta}_s < \eta < \overline{\eta}_s$ implies $(1 + 2g_s)/\eta \in (1, (1 + 2g_s)^2/4] \subset (1, 4]$, hence $\sqrt{(1 + 2g_s)/\eta} \in (1, 2]$ and so $1 - \sqrt{(1 + 2g_s)/\eta} \in [-1, 0)$, giving $\Delta(s,\eta) \leq -1$.

By (C.2), the corresponding slopes satisfy

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

Combine this ordering with the trust condition that the optimal point is $\mathsf{R}_s(\eta)$ rather than D_s

$$\Delta(s, \eta) \leq -\frac{\eta}{1-\eta}. \tag{C.3}$$

Since the right-hand side $-\eta/(1-\eta)$ is a scalar, the set of indices for which (C.3) holds is an initial segment: there exists $m(\eta) \in \{0, 1, \ldots, |\mathcal{S}|\}$ such that

$$\Delta(s^{(j)}, \eta) \leq -\frac{\eta}{1-\eta} \quad \Longleftrightarrow \quad j \leq m(\eta).$$

Equivalently, the optimizer selects $\mathsf{R}_{s^{(j)}}(\eta)$ for $j \leq m(\eta)$ and selects $\mathsf{D}_{s^{(j)}}$ for $j > m(\eta)$. By Step 1, this endpoint-by-endpoint selection is exactly the global maximizer.

Finally, the indifference slope $-\eta/(1-\eta)$ is strictly decreasing in η on $(0, 1)$. Fix $\eta_1 < \eta_2$. Then

$$-\frac{\eta_2}{1-\eta_2} < -\frac{\eta_1}{1-\eta_1}.$$

Hence the inequality $\Delta(s, \eta) \leq -\eta/(1-\eta)$ becomes (weakly) easier to satisfy as η increases, so the initial segment of indices satisfying it can only expand. Formally, by the definition of $m(\eta)$,

$$j \leq m(\eta_1) \;\Rightarrow\; \Delta(s^{(j)}, \eta_1) \leq -\frac{\eta_1}{1-\eta_1} \;\Rightarrow\; \Delta(s^{(j)}, \eta_2) \leq -\frac{\eta_2}{1-\eta_2} \;\Rightarrow\; j \leq m(\eta_2),$$

where the middle implication uses the fact that the optimizer's slope rule (C.3) is evaluated at the same η that determines $\Delta(\cdot, \eta)$. Therefore $m(\eta)$ is weakly increasing in η.

References

Ambrus, A. and S. Takahashi (2008): "Multi-sender cheap talk with restricted state spaces," *Theoretical Economics*, 3, 1–27.

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*.

Battaglini, M. (2002): "Multiple referrals and multidimensional cheap talk," *Econometrica*, 70, 1379–1401.

Bergemann, D. and S. Morris (2005): "Robust mechanism design," *Econometrica*, 73, 1771–1813.

——— (2019): "Information Design: A Unified Perspective," *Journal of Economic Literature*, 57, 3–30.

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*.

Crawford, V. P. and J. Sobel (1982): "Strategic information transmission," *Econometrica*, 1431–1451.

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

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.

Ghirardato, P., F. Maccheroni, and M. Marinacci (2004): "Differentiating Ambiguity and Ambiguity Attitude," *Journal of Economic Theory*, 118, 133–173.

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," *Available at SSRN 4723587*.