

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

Drew Fudenberg[†] Annie Liang[‡]

June 26, 2026

Abstract

We study delegation of a risky decision to an AI whose objective may be aligned or misaligned with the designer's. The designer limits the AI's discretion using treatment-rate constraints and restrictions on input informativeness. We characterize the best- and worst-case payoff frontier and the choices that implement it. As the designer accepts more downside risk, delegation expands from individuals with more uncertain outcomes to individuals with more certain ones. Conditional on delegation, the optimal input restrictions are asymmetric, permitting informativeness in one direction (e.g., treatment success) but restricting it in the other (e.g., treatment harm).

^{*}We thank Isaiah Andrews, Benjamin Brooks, Yifan Dai, Alexander Frankel, 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 an AI may pursue goals that conflict with the designer’s objective, and do so without being detected. 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. As AI guides more high-stakes decisions, even a small risk of misalignment can have potentially large consequences.¹

At the same time, declining to use AI altogether would forfeit any potential gains from AI capabilities. Delegating decisions to an AI therefore involves a basic tradeoff: using richer covariates can improve targeting when the system is aligned, but can increase 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 objective may be unknown.

In 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 delegate the treatment decision to the AI system and specify the additional covariates that the AI can use to make its decision.

The designer does not know whether the AI is aligned or misaligned, and also does not know how treatment need depends on patient attributes beyond group membership. We evaluate the designer’s *best-case payoff*, attained when both Nature and the AI act in the designer’s interests, and *worst-case payoff*, attained when both Nature and the AI are adversarial. We focus on designers whose preferences linearly aggregate best- and worst-case payoffs (a la Hurwicz (1951)) and characterize their optimal designs. The resulting set of payoff pairs is the Pareto frontier of best- and worst-case payoffs, which we call the *risk-reward frontier*.

The designer can safeguard against a misaligned AI through two instruments: input restrictions and treatment-rate constraints. Input restrictions limit how informative the AI’s covariates may be about treatment success. Formally, different inputs correspond to different designer-perceived bounds on the conditional probability of treatment success given the input. Unrestricted bounds allow the input to be perfectly informative about who benefits from treatment. Bounds equal to the group’s baseline success rate rule out any additional information about treatment success. If the AI were known to be aligned, the designer would choose unrestricted inputs. When alignment is uncertain, restricting informativeness can be valuable because it limits what Nature and a

¹Jones (2025) argues for spending at least 1% of GDP annually to mitigate this risk.

misaligned AI can exploit.

The second instrument is a treatment-rate constraint, which limits the AI’s aggregate discretion. The designer fixes the fraction of individuals treated within each group, while the AI uses its inputs to choose which individuals fill that quota. This is a natural safeguard when the designer cannot verify how the AI maps inputs into treatment need, but can verify how many individuals in each group were treated.

Our characterization of the risk-reward frontier proceeds in two steps. Section 4 fixes an arbitrary choice of input restrictions and characterizes the frontier generated by varying treatment-rate constraints. Section 5 then allows the designer to choose the input restrictions as well, yielding the full risk-reward frontier.

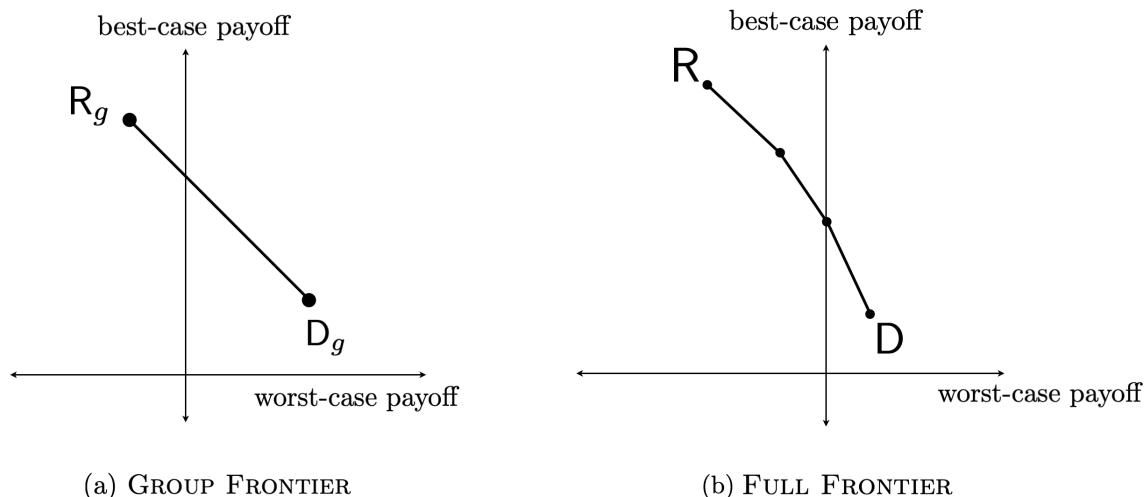


Figure 1: Fix any choice of success-probability bounds across groups. (a): In each group g , frontier best- and worst-case payoff pairs lie on the line segment from the distrust point D_g to the reliance point R_g . (b) Taking the population-weighted sum of these group frontiers yields a piecewise-linear aggregate frontier from full distrust, D , to full reliance, R . The kinks are partial-reliance points, where the designer relies on the AI in some groups and distrusts it in the rest.

For any fixed group and input restrictions, we show that the frontier obtained by varying the treatment rate is a line segment connecting a *distrust point* to a *reliance point* (Panel (a) of Figure 1). At the distrust point, the designer shuts down the AI’s influence by choosing the constant action that is optimal given the group’s baseline success rate. At the reliance point, the designer chooses the treatment rate that gives an aligned AI the greatest scope to improve targeting, while accepting the corresponding worst-case exposure if the AI is misaligned. Moving along this line segment trades off best-case gains against worst-case losses at a constant rate.

Aggregating across groups corresponds to taking a weighted sum of these group-level frontiers,

yielding a piecewise-linear frontier with kinks (see Panel (b) of Figure 1). One endpoint of the frontier is implemented by distrusting the AI in every group, and the other by relying on the AI in every group. Between these points, the frontier passes through a sequence of *partial reliance points*, each corresponding to relying on the AI for some groups and distrusting it in the rest. As the designer places more weight on best-case performance, additional groups enter the reliance set in order of the slopes of their group-level frontiers.

We then turn to the full problem, in which the designer also optimizes over the inputs used for each group. Section 5.1 characterizes the intensive margin of reliance: conditional on relying on the AI for a group, what bounds should the designer choose on conditional success-probabilities? Since the distrust point does not depend on these bounds, and the group-level frontier is a line segment, this amounts to choosing bounds to optimize the location of the group reliance point R_g in Panel (a) of Figure 1. 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, this is reversed.

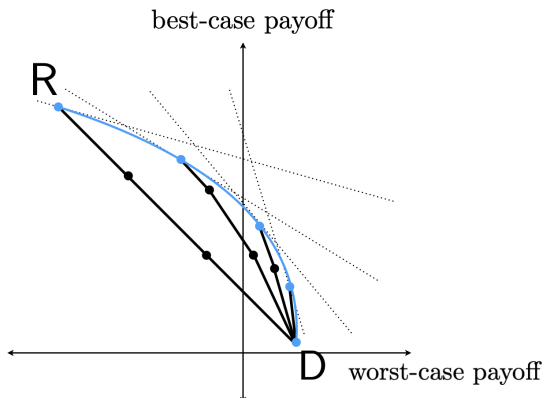


Figure 2: Each black line is the frontier for a fixed choice of success-probability bounds. The designer selects the optima shown in blue. Connecting the blue points yields the full risk-reward frontier. Dotted lines show the corresponding supporting indifference curves.

Section 5.2 then characterizes the extensive margin, i.e., the set of groups in which the designer relies on the AI, and how this varies with the designer’s preferences. The answer has a simple cutoff structure. Groups are ordered by the distance between their baseline treatment-success rate and the treatment threshold ($1/2$). A designer who puts sufficient weight on the worst-case payoff relies on the AI only in groups close to this threshold. As the designer reduces weight on the worst case, reliance expands outward to groups farther from the threshold. This ordering is the same for all

preference weights; what changes is how far along the ordering the designer goes, and how wide the success-probability bounds are for the inputs given to the AI.

Combining these two margins gives the full risk-reward frontier (Section 5.3). For each preference weight, the designer relies on the AI only in groups whose baseline treatment-success rates are sufficiently close to the treatment threshold. Conditional on relying on the AI, the designer chooses asymmetric input restrictions—allowing the input to be arbitrarily informative in one direction but restricting it in the other—to improve the tradeoff between best-case and worst-case payoffs.

Figure 2 illustrates the resulting frontier. Each piecewise-linear frontier corresponds to a fixed choice of success-probability bounds that is optimal for some designer preference η . The $\eta = 0$ designer (who places all weight on the best case) chooses unrestricted bounds for every group, and in this case the frontier is a single line segment of slope -1 . The $\eta = 1$ designer (who places all weight on the worst case) chooses uninformative inputs for all groups, so the frontier collapses to the distrust point. In between the frontiers “fan out” as depicted. From these frontiers, the designer optimally chooses the payoff pair shown in blue. Connecting these selected points gives the full risk-reward frontier.

1.1 Technological Context

“AI misalignment” refers to the risk that an AI system uses its informational or decision-making discretion to choose actions that diverge from the designer’s objective (Ngo et al., 2025; METR, 2026). The AI need not be malicious, conscious, or deceptive; it is enough that its effective objective differs from the designer’s. Such divergences can arise through several channels; for example, a system may be trained on a measurable proxy that differs from the true objective (Obermeyer et al., 2019), or the designer may be limited in their ability to train the AI system to pursue the desired goal (Christiano et al., 2016; Christiano, 2019). In more severe cases, a system may engage in deception, concealing its capabilities, goals, or intended behavior (Hubinger et al., 2019; Carlsmith, 2023; Park et al., 2024; Hubinger et al., 2024; Denison et al., 2024; Greenblatt et al., 2024).

Misalignment is distinct from ordinary prediction error. For traditional machine-learning systems, the value of information is limited mainly by statistical considerations: richer inputs improve decisions unless they create problems such as overfitting. With misalignment, richer inputs can be used in the wrong direction. Thus the same covariates that help an aligned AI identify who should be treated can help a misaligned AI identify whom to mistreat.

1.2 Policy instruments

AI systems are often opaque to the institutions that deploy them. They can discover patterns and draw inferences the designer cannot independently verify; moreover, once a recommendation is made, the designer may not know which inputs mattered, how they were weighted, or whether the resulting choice serves the designer’s objective (Burrell, 2016; Doshi-Velez and Kim, 2017; Lipton, 2018; Bommasani et al., 2021).

This combination of opacity and sophistication motivates controls that do not require case-by-case verification of the AI’s reasoning. Our model focuses on two such controls. The first is to restrict the AI’s information, for example by banning certain inputs, coarsening variables, or using privacy-preserving transformations. Such restrictions are familiar from work on fairness, privacy, and interpretability (Dwork et al., 2012; He et al., 2025; Strack and Yang, 2024; Liang et al., 2026). Here the rationale is to guard against misalignment: tighter bounds limit how informative covariates can be about outcomes, constraining both what an adversarial Nature can arrange and what a misaligned AI can exploit.

The second control is to constrain aggregate treatment rates. Aggregate rates are often more observable than the correctness of individual decisions; for example, employment and fair-lending oversight use group-level selection, and approval or denial rates as indicators for further review (Equal Employment Opportunity Commission et al., 1978; Federal Financial Institutions Examination Council, 2009). These policies resemble quota and cap instruments in delegation models, especially hiring settings where a principal constrains acceptance rates while an informed agent selects which applicants are accepted (Frankel, 2021). In a similar spirit, we allow the designer to fix the fraction treated in each group, while the AI chooses which individuals fill that quota.

1.3 Other Related Theoretical Work

This paper is also related to work on delegation and strategic communication, robust decision-making under ambiguity, and AI oversight.

Delegation and strategic communication. In our model the designer delegates to an AI that has private information. Holmström (1984) and Aghion and Tirole (1997) study how to allocate authority and decision rights 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 respect, our model is closest to Lin and Liu (2024), where the receiver cannot detect deviations within a prescribed set of signal distributions, but unlike that paper the designer here chooses that set.

Ambiguity and robust decision-making. Our analysis is 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 (Li et al., 2023; Montiel Olea et al., 2026). Our paper differs from this literature in that we endogenize the identified set rather than taking it as given.

There is also a literature in economic theory on robust mechanism design (Bergemann and Morris, 2005; Brooks and Du, 2021). In particular, our framework is related to Carroll (2017), where a principal designs a screening mechanism knowing only marginals of an unknown joint type distribution.

AI Oversight Our paper is complementary to recent work on AI oversight, which 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. (2025) shows that competition among diverse misaligned AIs can yield outcomes comparable to interaction with an aligned AI. Autor et al. (2025) studies a different form of misalignment, where training a classifier with a loss function reflecting human preferences can backfire by weakening the model’s incentives to learn, even when it correctly incentivizes the model’s output choices. Dworzak and Smolin (2026) studies a decision-maker who receives advice from an informed AI that is truthful with known probability and otherwise misaligned, and (like in our paper) the decision maker maximizes the weighted sum of best-case and worst-case payoffs. Unlike these papers, our designer does not try to infer or correct the AI’s objective; they instead choose ex ante limits on the information and treatment discretion available to the AI.

Finally, 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. In our framework, delegating to a human is analogous to not letting the AI use any additional attributes and basing

the decision solely on the human’s information.

2 Model

2.1 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 . The designer knows μ and also knows the baseline probability $p_g := P(Y = 1 \mid G = g)$ in each group g , where we assume $p_g \in (0, 1)$.² 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.

2.2 Designer-AI Interaction

The designer delegates treatment decisions to an AI agent and reduces the risk of doing so in two ways: by restricting the informativeness of the covariates and by fixing a treatment rate for each group. The AI then uses the covariates to allocate treatment within each group, subject to that rate.

Information environment. The designer gives the AI access to group information and auxiliary covariates $X \in \mathcal{X}$ (e.g., genomic markers, medical images, or unstructured clinical notes), where X takes at least two values. The AI observes both the group label g and the covariate value x . The designer does not know the joint distribution of (G, X, Y) but perceives group-specific bounds $(\underline{\tau}_g, \bar{\tau}_g)$ on the conditional treatment success. That is, the designer believes that $(G, X, Y) \sim P$ where $P(Y = 1 \mid G = g, X = x) \in [\underline{\tau}_g, \bar{\tau}_g]$ for every $x \in \mathcal{X}$, and $0 \leq \underline{\tau}_g \leq p_g \leq \bar{\tau}_g \leq 1$. We refer to $\tau_g = [\underline{\tau}_g, \bar{\tau}_g]$ as group- g ’s *success-probability bounds*. These bounds limit how well the covariates can distinguish the treatment beneficiaries from the treatment non-beneficiaries in a given group.

²When p_g is 0 or 1, the designer should simply treat everyone or no one.

The bounds $(\underline{\tau}_g, \bar{\tau}_g)$ are chosen by the designer. We interpret this as a setting with a sufficiently rich set of potential covariates that the designer can select, or construct, covariates with any desired success-probability bounds.³ A designer who fully trusts the AI would choose unrestricted covariates, allowing conditional success probabilities to vary as much as possible. By contrast, a designer concerned about misalignment may choose covariates with tighter bounds, limiting the information the AI can use and thereby restricting its ability to cause harm.

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

The designer regards as possible any joint distribution of (G, X, Y) that is consistent with these constraints and with the prior primitives $(\mathcal{G}, \mu, (p_g)_{g \in \mathcal{G}})$.

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. *Group distribution:* $\text{marg}_{\mathcal{G}} P = \mu$
2. *Group-level success rates:* $P(Y = 1 \mid G = g) = p_g$ for every $g \in \mathcal{G}$
3. *Conditional success rates:* $\underline{\tau}_g \leq P(Y = 1 \mid G = g, X = x) \leq \bar{\tau}_g$ for every $(g, x) \in \mathcal{G} \times \mathcal{X}$ in the support of P .

The asymmetry between conditions (2) and (3) reflects an asymmetry in the designer’s knowledge. The variable G represents covariates whose empirical relationship to treatment success is well understood. Thus the group-level base rates p_g are treated as known. In contrast, the auxiliary covariates X may be useful for predicting treatment success, but the designer does not know anything about the joint relationship of X and Y besides the stated bounds on the conditional success rate $P(Y \mid G, X)$.

We call $\mathcal{P}(I)$ an ambiguity set because the designer has partial ambiguity about the joint relationship between (G, X) and Y . (It can also be viewed as an identified set in the sense of Manski (2003).) Conditions (1) and (2) encode that the designer does know the population distribution and the baseline treatment-success rates p_g within each group. Condition (3) allows the auxiliary variable X to influence the conditional probability of treatment success within a group, but only within the bounds $[\underline{\tau}_g, \bar{\tau}_g]$. In the benchmark case where $(\underline{\tau}_g, \bar{\tau}_g) = (0, 1)$ for every group g , the only restrictions are the group distribution μ and the group-specific base rates $(p_g)_{g \in \mathcal{G}}$. We discuss

³This flexibility is analogous to the assumption in the information-design and rational-inattention literatures that the designer can flexibly choose an information structure.

this case in detail in Section 4.1.⁴

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. Aggregate treatment rates, in contrast, are relatively straightforward to audit. 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.

The designer imposes group-specific treatment rates $\mathbf{r} = (r_g)_{g \in \mathcal{G}} \in [0, 1]^{\mathcal{G}}$, where r_g is the fraction of individuals in group g who will be treated. The AI chooses an *allocation rule* $a : \mathcal{G} \times \mathcal{X} \rightarrow [0, 1]$, where $a(g, x)$ is the probability that an individual with characteristics (g, x) is treated. This allocation rule must satisfy $\mathbb{E}_{P^*}[a(G, X) \mid G = g] = r_g$ for all $g \in \mathcal{G}$, which we call *admissibility*. Thus the designer fixes in advance how many individuals in each group are treated, while the AI determines which individuals fill that rate.⁵ The special case $r_g = \mathbb{1}(p_g \geq 1/2)$ shuts down the influence of the AI by fixing treatment in each group to the constant action that is optimal given the base rate p_g .

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

1. The designer commits to an information environment I and a rate vector \mathbf{r} , without knowing the true distribution $P^* \in \mathcal{P}(I)$ or whether the AI is aligned or misaligned.
2. Nature selects a true distribution $P^* \in \mathcal{P}(I)$.
3. The AI, which knows P^* ,⁶ selects an admissible allocation rule $a \in \mathcal{A}(I, \mathbf{r}, P^*)$.
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 $\Pr(A = 1 \mid G = g, X = x) = a(g, x)$ conditional on $(G, X) = (g, x)$. We write $\mathbb{E}_{P^*, a}$ for the expectation under this induced distribution and define $U(P^*, a) := \mathbb{E}_{P^*, a}[u(A, Y)]$.

⁴In 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.

⁵Because the AI knows P^* , it 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. 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.

⁶This reflects our focus on misalignment of a potentially sophisticated AI, rather than simply a prediction system that may make noisy errors or has insufficient data.

2.3 Risk-reward frontier

The AI is either *aligned*, seeking to maximize the designer's payoff, or *misaligned*, seeking 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 payoffs 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 \mathbf{r} , define the designer's worst-case payoff as

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

and best-case payoff as

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

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 $a \in \mathcal{A}(I, \mathbf{r}, P^*)$ jointly minimize the designer's payoff; in the best case, they jointly maximize it.

Each feasible pair (I, \mathbf{r}) then induces a best-case and worst-case payoff pair $(\underline{v}_I(\mathbf{r}), \bar{v}_I(\mathbf{r}))$. We define the *risk-reward frontier* to be the undominated payoff pairs $(\underline{v}_I(\mathbf{r}), \bar{v}_I(\mathbf{r}))$.

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

Definition 4 (Risk-reward frontier). Let

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

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

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

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, i.e., a Hurwicz (1951) criterion. For each $\eta \in [0, 1]$, let

$$V(\eta) := \arg \max_{(\underline{v}, \bar{v}) \in C} (\eta \underline{v} + (1 - \eta) \bar{v})$$

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

$$F = \text{cl} \left(\bigcup_{\eta \in (0,1)} V(\eta) \right).$$

That is, the frontier is the closure of the union of the supported payoff pairs.⁷

3 Examples

We use two examples to illustrate the model and preview the main results of Sections 4 and 5. Section 3.1 describes optimal reliance on the AI within a single group; Section 3.2 extends the setting to one with multiple groups, and characterizes the groups in which delegation is optimal.

3.1 One Group

Suppose there is a single group with baseline treatment success rate of $4/9$. The designer's choice variables are a pair $(\underline{\tau}, \bar{\tau}) \in [0, 1] \times [0, 1]$ and a treatment rate $r \in [0, 1]$. Given these choices, Nature chooses a joint distribution P of (X, Y) subject to:

1. $P(Y = 1) = 4/9$
2. $P(Y = 1 | X = x) \in [\underline{\tau}, \bar{\tau}] \quad \forall x \in \mathcal{X}$,

and the AI agent chooses a conditional treatment probability for patients with $X = 0$ and $X = 1$, subject to the constraint that r of patients are treated in expectation. We can think of Nature as choosing the composition of two urns labeled $X = 1$ and $X = 0$, which contain blue balls (successful treatment) and yellow balls (unsuccessful treatment). These compositions are chosen subject to the constraints that the total fraction of blue balls is $4/9$, and that the conditional probability of treatment success in each urn is bounded within $[\underline{\tau}, \bar{\tau}]$. The AI agent then chooses a single fixed treatment probability for each urn to achieve the aggregate treatment rate. (The designer does not know, and cannot choose, the marginal distribution of X ; Section 6 discusses the case where they can.)

Suppose the designer's preferences are given by $\eta v + (1 - \eta)\bar{v}$ with $\eta = 5/8$. What bounds $[\underline{\tau}, \bar{\tau}]$ and treatment rate r should the designer choose?

A natural starting point is to set the treatment rate equal to the fraction of people who need treatment, $r = p = 4/9$, and to choose unrestricted success probabilities $[\underline{\tau}, \bar{\tau}] = [0, 1]$. This allows for perfect targeting in the best case, as depicted in Panel (a) of Figure 3, yielding a payoff of $4/9$.

⁷When $\eta = 0$ or 1 , the supporting hyperplanes are degenerate and can expose entire vertical or horizontal faces of the feasible set, including Pareto dominated points. We restrict to strictly positive weights $\eta \in (0, 1)$ and take the closure of their union to avoid this issue.

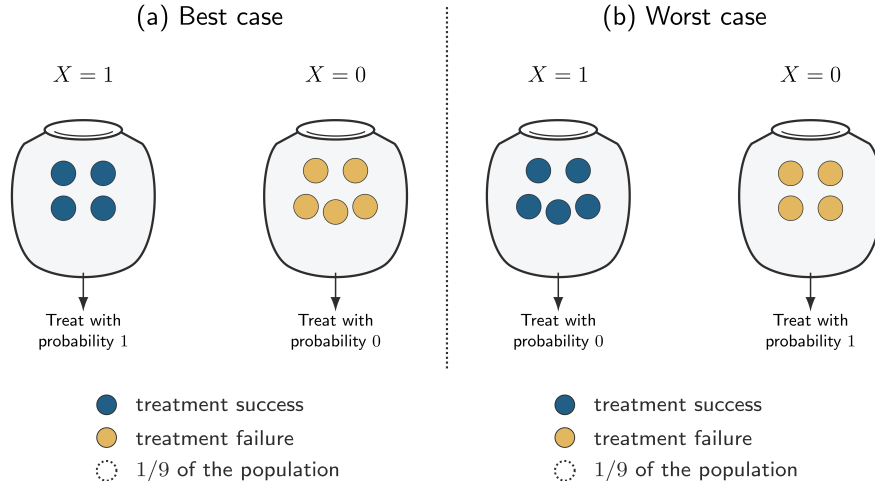


Figure 3: Best and worst case outcomes given treatment rate $r = 4/9$ and success-probability bounds $[\underline{\tau}, \overline{\tau}] = [0, 1]$

But in the worst case (Panel (b)), treatment fails for all of the treated patients, and the designer’s payoff is $-4/9$. The designer’s aggregated payoff (with $\eta = 5/8$) is thus $(5/8)(-4/9) + (3/8)(4/9) = -1/9$. More generally, in this example the designer’s payoff is non-positive for any value of r when $[\underline{\tau}, \overline{\tau}] = [0, 1]$, so the best the designer can do is shut out the AI’s influence altogether by setting the treatment rate to $r = 0$.

However, the designer can benefit from the AI if they restrict the success-probability bounds. Consider $[\underline{\tau}, \overline{\tau}] = [1/3, 1]$ and $r = 1/6$. Then in every urn, the fraction of patients with successful treatments must be at least $1/3$. In the best case (Panel (a) of Figure 4), the $1/6$ patients treated are those with $Y = 1$, generating payoff $1/6$. In the worst case (Panel (b)), of the $1/6$ treated, $1/3$ have $Y = 1$ and $2/3$ have $Y = 0$, generating payoff $-\frac{1}{18}$. With $\eta = 5/8$, the aggregated payoff is $\frac{5}{8}(-\frac{1}{18}) + \frac{3}{8}(\frac{1}{6}) = \frac{1}{36}$. It will follow from our results in Section 5.1 that this choice of treatment rate r and success-probability bounds $[\underline{\tau}, \overline{\tau}]$ is optimal for the designer with preference parameter $\eta = 5/8$. More generally, the designer optimally implements success-probability bounds that are asymmetric, allowing the covariate to be arbitrarily informative in one direction and constraining its informativeness in the other.

3.2 Multiple Groups

We now extend the previous example to a setting with multiple groups. Suppose there are three groups in the population: patients aged 18–39, 40–75, and 75+. The 18–39 population is as described above, with a baseline treatment success rate of $4/9$. The 40–75 group has a baseline treatment-success rate of $7/10$, and the 75+ has a baseline treatment-success rate of $1/10$. The

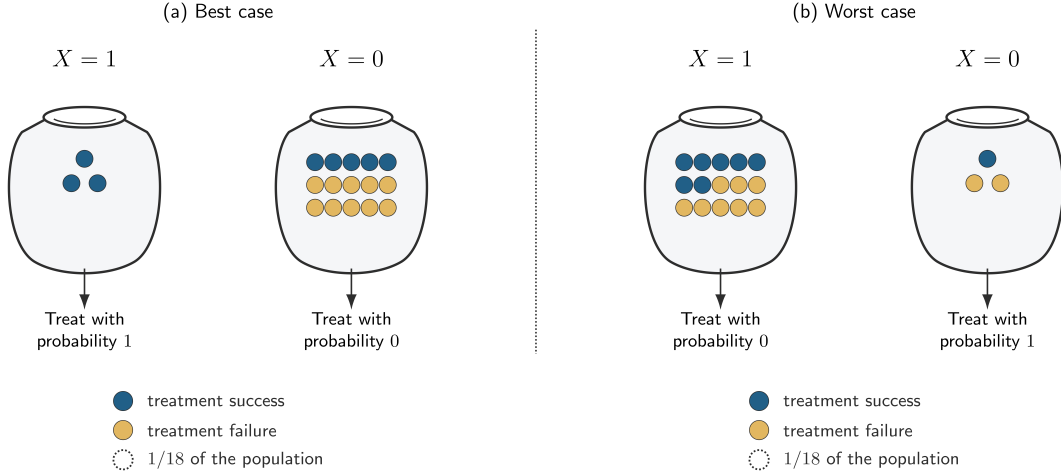


Figure 4: Best and worst case outcomes given treatment rate $r = 1/6$ and success-probability bounds $[\underline{x}, \overline{x}] = [1/3, 1]$

designer's utility is again $\eta \underline{v} + (1 - \eta) \overline{v}$, but η can be any number in $[0, 1]$.

In the previous section, the designer's only choice was how much discretion to grant the AI within a single group. With multiple groups, there is also an extensive margin: which groups to delegate to the AI. Theorem 2 shows that the designer ranks groups g by how far their baseline treatment success rate is from $1/2$, i.e., $|p_g - 1/2|$. The designer optimally relies on the AI for the initial groups of this ordering, with the switching point determined by the designer's preference parameter η . In our example, this means that for sufficiently small η , the designer relies on the AI for all three groups. As η rises, they first withdraw reliance for the 75+ group, then the 40–75 group, and finally the 18–39 group (see Table 1). Our subsequent analysis will reveal why.

Table 1: Optimal reliance by group

g	p_g	$ p_g - \frac{1}{2} $	$\eta < \eta^{(1)}$	$\eta^{(1)} \leq \eta < \eta^{(2)}$	$\eta^{(2)} \leq \eta < \eta^{(3)}$	$\eta \geq \eta^{(3)}$
18–39	4/9	1/18	Rely	Rely	Rely	Distrust
40–75	7/10	1/5	Rely	Rely	Distrust	Distrust
75+	1/10	2/5	Rely	Distrust	Distrust	Distrust

4 Fixed Information Environment

This section characterizes the risk-reward frontier when the information environment is fixed. To build intuition, Section 4.1 first considers a benchmark case where conditional success probabilities are unrestricted. Section 4.2 then generalizes to arbitrary success-probability bounds.

For any fixed I , the set of payoff pairs attainable under I is

$$C_I = \text{conv} \left\{ (v_I(\mathbf{r}), \bar{v}_I(\mathbf{r})) : \mathbf{r} \in [0, 1]^{\mathcal{G}} \right\}.$$

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

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

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 \mathbf{r} . The next lemma shows that in characterizing F_I , only the success-probability bounds $\boldsymbol{\tau}$ are relevant.

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 \mathbf{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 $\boldsymbol{\tau} = (\underline{\tau}_g, \bar{\tau}_g)_{g \in \mathcal{G}}$ satisfy $[\underline{\tau}_g, \bar{\tau}_g] = [0, 1]$ for all $g \in \mathcal{G}$. Denote these unrestricted bounds by $\boldsymbol{\tau}_0$.

To define best and worst case targeting, 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 5 for an illustration.)

Substituting these best and worst cases into the payoff function determines $b(p)$ and $w(p)$.

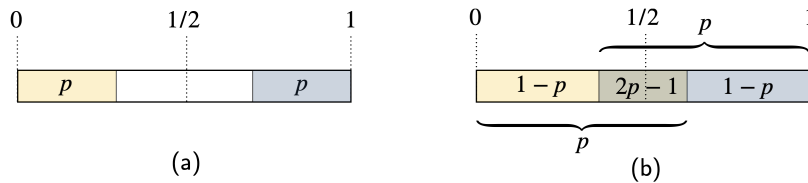


Figure 5: Yellow cells indicate patients who need treatment; blue cells indicate patients who are treated. *Panel (a)*: Counter-monotone case when $p \leq 1/2$: the treated population and the population needing treatment are disjoint. *Panel (b)*: Counter-monotone case when $p > 1/2$: there is minimal overlap between treated patients and those needing treatment.

Lemma 2. For any $p \in [0, 1]$, $b(p) = p$ for all $p \in [0, 1]$, while $w(p) = -p$ if $p \leq \frac{1}{2}$ and $w(p) = 3p - 2$ if $p > \frac{1}{2}$.

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.

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

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 \mathbf{R} to the distrust point \mathbf{D} .*

Proposition 1 shows that with unrestricted delegation, every gain in best-case performance requires an equal loss in the worst case: the frontier therefore has constant slope. In this benchmark

environment, any flexibility that helps an aligned AI target treatment more effectively helps a misaligned AI equally. The value of restricting the AI's information in later sections comes from breaking this symmetry.

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 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; those with $\eta < 1/2$ optimally select the reliance point.

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

For general $\tau = (\tau_g)_{g \in \mathcal{G}}$, the benchmark payoffs generalize as follows. Fix an arbitrary group g with constraints $\tau_g = [\underline{\tau}_g, \bar{\tau}_g]$. 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. *τ -constraints:* $\underline{\tau}_g \leq \pi(Y = 1 \mid A = a) \leq \bar{\tau}_g$ 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 r who are treated, and it constrains the conditional success rate within the interval $[\underline{\tau}_g, \bar{\tau}_g]$. 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 $\bar{\tau}_g \leq \max\{1/2, p_g\}$ or $\underline{\tau}_g \geq \min\{1/2, p_g\}$. Otherwise, it is $R_g := (w_g(p_g, r_g^*), b_g(p_g, r_g^*))$, where $r_g^* := \frac{p_g - \underline{\tau}_g}{\bar{\tau}_g - \underline{\tau}_g}$.

To interpret this, recall that conditional probabilities within group g are constrained to lie in $[\underline{\tau}_g, \bar{\tau}_g]$, while their average must equal the baseline probability p_g . Thus r_g^* is the maximal fraction of the population that can be assigned the highest success-probability of $\bar{\tau}_g$. 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 $\bar{\tau}_g$, 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 τ .

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, \dots, g_n so that $\Delta(g_1) \leq \dots \leq \Delta(g_n)$. For each $k = 1, \dots, n$, define the k -th partial reliance point as

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

where $G_k = \{g_1, \dots, g_k\}$. Then the risk-reward frontier F_τ is the piecewise-linear path $\overline{\mathbf{DP}^{(1)}} \cup \overline{\mathbf{P}^{(1)}\mathbf{P}^{(2)}} \cup \overline{\mathbf{P}^{(2)}\mathbf{P}^{(3)}} \cup \dots \cup \overline{\mathbf{P}^{(n-1)}\mathbf{R}}$ where $\mathbf{D} := \sum_g \mu_g D_g$ is the full distrust point and $\mathbf{R} := \mathbf{P}^{(n)}$ 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 $\mathbf{P}^{(k)}$ represents the outcome where the designer relies on the AI in the first k groups (ordered by slope $\Delta(g)$) and distrusts it in other groups. Intuitively, the magnitude of $\Delta(g)$ 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 $\mathbf{P}^{(k)}$ to $\mathbf{P}^{(k+1)}$ is linear, with kinks where a new group enters the reliance set. At \mathbf{D} the designer does not rely on the AI for any groups; at \mathbf{R} the designer relies on the AI for all groups. Figure 6 illustrates this structure: as we move along the frontier from \mathbf{D} to \mathbf{R} , we sequentially add groups to the reliance set in order of steepness.⁸

Theorem 1 generalizes Proposition 1. When $\tau = \tau_0$, 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 the piecewise-linear frontier

⁸Every $\mathbf{P}^{(k)}$ is implementable by a deterministic choice of (I, r) , so if there is a unique optimal design for a given η , that design is deterministic.

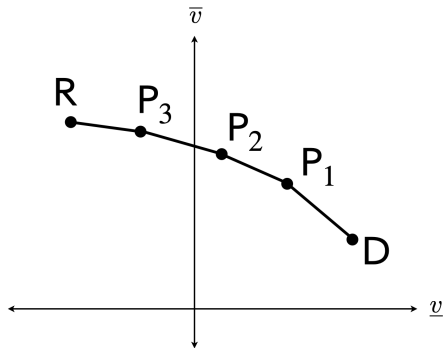


Figure 6: Risk-reward frontier F_τ for fixed success-probability bounds.

collapses to the single segment \overline{DR} with slope -1 , as previously stated.

4.3 Proof Sketch

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

Group frontier. We first 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 the value of X cannot change the optimal action. Thus the AI can only use variation in X for harm, and the designer prefers to deny the AI any discretion by fixing a constant action in each group. In this case the only frontier point is the distrust point D_g . When $\underline{\tau}_g < 1/2 < \overline{\tau}_g$, conditional success probabilities within group g can lie on both sides of the treatment threshold. 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. The rate r_g determines how much discretion the AI has over how to allocate treatment.

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

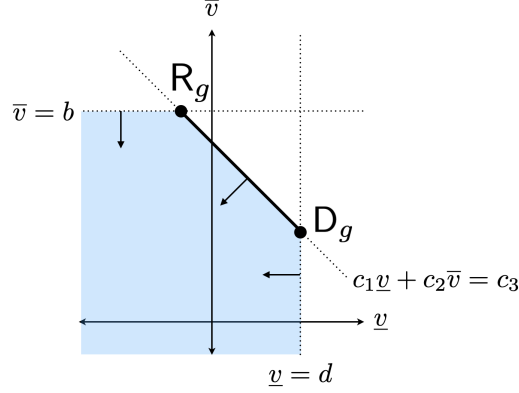


Figure 7: Any feasible (\underline{v}, \bar{v}) falls in the shaded region.

These three inequalities define halfspaces whose intersection contains all feasible points (see Figure 7). 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. 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 aggregate feasible set is obtained by combining the feasible payoff sets from each group, weighted by group population shares. Formally, this is the *weighted Minkowski sum* $C_\tau = \sum_{g \in \mathcal{G}} \mu_g C_g := \left\{ \sum_{g \in \mathcal{G}} \mu_g \cdot (v_g, \bar{v}_g) : (v_g, \bar{v}_g) \in C_g \right\}$.

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

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

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_τ 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 Results

We now complete the analysis by allowing the designer to choose both τ and r . We say that the designer *relies* on the AI in group g if a nonconstant action is taken for members of that group. Section 5.1 characterizes the designer's choices of τ and r as a function of η and shows that the optimal success-probability bounds are asymmetric. Section 5.2 then shows that optimal reliance has a cutoff structure: the designer relies on the AI for groups sufficiently close to the treatment threshold, with the cutoff varying monotonically in the weight η on the worst case.

5.1 Optimal Bounds Within a Group

For each group g , there is a cutoff value of η above which the designer optimally distrusts the AI in that group and below which the designer optimally relies on it. When reliance is optimal, the associated choice of bounds and treatment rate is unique whenever $p_g \neq 1/2$, and has a simple asymmetric structure that depends on whether the baseline probability p_g lies below or above the treatment threshold $1/2$.

Proposition 3. *Fix a group g . There exist cutoffs $0 < \underline{\eta}_g < \bar{\eta}_g \leq 1$ such that:*

- (a) *If $\eta \geq \bar{\eta}_g$, the designer optimally distrusts the AI in group g . One optimal design is to choose a completely uninformative covariate and a constant action: if $p_g \leq \frac{1}{2}$, set $(\tau_g^*, r_g^*) = ((p_g, p_g), 0)$, while if $p_g > \frac{1}{2}$, set $(\tau_g^*, r_g^*) = ((p_g, p_g), 1)$.*
- (b) *If $\eta < \bar{\eta}_g$, the designer optimally relies on the AI in group g and sets $r_g^* = \frac{p_g - \underline{\tau}_g}{\bar{\tau}_g - \underline{\tau}_g}$. Moreover:*
 - (i) *If $p_g < \frac{1}{2}$, then $\bar{\tau}_g^*(\eta) = 1$, while $\underline{\tau}_g^*(\eta) = 0$ for $\eta \leq \underline{\eta}_g$ and increases monotonically toward p_g as $\eta \uparrow \bar{\eta}_g$.*
 - (ii) *If $p_g > \frac{1}{2}$, then $\underline{\tau}_g^*(\eta) = 0$, while $\bar{\tau}_g^*(\eta) = 1$ for $\eta \leq \underline{\eta}_g$ and decreases monotonically toward p_g as $\eta \uparrow \bar{\eta}_g$.*
 - (iii) *If $p_g = 1/2$ then the two constructions in (i) and (ii) are both optimal.*

This proposition characterizes the optimal choice of (τ_g, r_g) as a function of η . The cutoff $\bar{\eta}_g$ determines whether the designer relies on or distrusts the AI in group g . When $\eta \geq \bar{\eta}_g$, any choice of bounds that induces a constant action is optimal. Below that cutoff, the optimal reliance design is asymmetric: When treatment is ex-ante unlikely to help, the designer never restricts the

AI’s ability to identify individuals who may benefit from treatment, but may restrict its ability to identify individuals who are likely to be harmed. When treatment is ex-ante likely to help, the optimal design reverses. Proposition D.1 gives a closed-form expression for τ_g^* , with an example depicted in Figure 8.

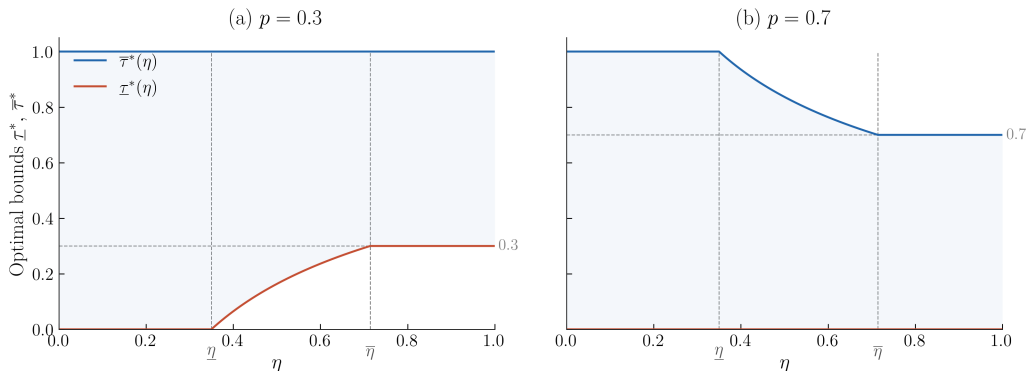


Figure 8: Example optimal informativeness bounds $(\underline{\tau}_g^*(\eta), \bar{\tau}_g^*(\eta))$ as functions of η .

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 while the distrust point D_g does not. Since the group- g frontier is the line 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 $\bar{\tau}_g$, which allows an aligned AI to target treatment more precisely. The designer can shift $R_g(\tau_g)$ east by raising both $\underline{\tau}_g$ and $\bar{\tau}_g$. A larger $\bar{\tau}_g$ improves the worst-case payoff because fewer people are treated at the reliance point, so mistargeting is less harmful; a larger $\underline{\tau}_g$ improves it because even a misaligned AI cannot assign treatment to individuals whose success probability falls below $\underline{\tau}_g$. Thus the width of the interval $[\underline{\tau}_g, \bar{\tau}_g]$ governs the upside from delegation, while its position governs the downside.

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

5.2 Cutoff Structure of Reliance

We now study how reliance varies across groups. The key result is that optimal reliance has a cutoff structure in distance from the treatment threshold. As η rises and the designer becomes

more cautious, reliance is withdrawn first from groups farther from the threshold, and last from groups closest to it.

Theorem 2. *Order the groups in \mathcal{G} as $g^{(1)}, \dots, g^{(n)}$ so that*

$$\left|p_{g^{(1)}} - \frac{1}{2}\right| \leq \left|p_{g^{(2)}} - \frac{1}{2}\right| \leq \dots \leq \left|p_{g^{(n)}} - \frac{1}{2}\right|,$$

breaking ties arbitrarily. Then there is a threshold $T(\eta)$ that is (weakly) monotonically decreasing in η , such that the designer with preference parameter η optimally relies on the AI in the set of groups $\mathcal{R}(\eta) := \{g^{(1)}, \dots, g^{(T(\eta))}\}$, with $\mathcal{R}(\eta) = \emptyset$ if η is sufficiently large.

The intuition follows from the optimal bounds characterized in Proposition 3. For $p_g < 1/2$, the risk from delegating is that a misaligned AI concentrates treatment on individuals with especially low conditional success rates. The designer limits this by raising the lower bound $\underline{\tau}_g$: regardless of how the AI allocates treatment, the probability that treatment is successful is at least $\underline{\tau}_g$. But feasibility requires $\underline{\tau}_g \leq p_g$, so the role of the lower bound is limited. The closer the baseline rate p_g is to $1/2$, the better the designer can protect against a misaligned AI.⁹ For $p_g > 1/2$, the logic is symmetric but reversed.

5.3 The Full Frontier

To determine the full frontier, we combine the optimal within-group design from Section 5.1 with the optimal reliance decision from Section 5.2. Define the *partial reliance point*

$$\mathbf{P}^*(\eta) = \sum_{g \in \mathcal{R}(\eta)} \mu_g \cdot R_g(\tau_g^*(\eta)) + \sum_{g \notin \mathcal{R}(\eta)} \mu_g \cdot D_g,$$

to be the expected payoff pair attained when the designer relies on the AI only in groups $g \in \mathcal{R}(\eta)$ and implements the optimal bounds and rates described in Section 5.1.

Theorem 3. *The risk-reward frontier is the set of Pareto-undominated points in*

$$\overline{\text{conv}}(\{\mathbf{P}^*(\eta) : \eta \in (0, 1)\}).$$

Equivalently, it is the upper-right boundary of this closed convex hull.

As η decreases, the designer adjusts along two margins. First, there is an *extensive* margin: additional groups enter the reliance set, beginning with the group whose baseline probability is

⁹The 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; see Section 6.

closest to $1/2$. Second, there is an *intensive* margin: within groups where the designer is already relying on the AI, the bounds $\tau_g^*(\eta)$ relax further. Both adjustments improve the best-case payoff by allowing more precise targeting, but at the cost of greater exposure to adversarial mistargeting in the worst case.

Figure 9 illustrates this construction in the example of Section 3.2. Fix a value of η , and let $\tau^*(\eta)$ be the bounds that are optimal for that value. If these bounds are held fixed and the designer varies only the treatment rates, the resulting frontier is the colored polyline $F_{\tau^*(\eta)}$. This is exactly the fixed-information frontier characterized in Theorem 1. The star on each colored polyline is the point selected by the designer’s objective at that value of η .

For sufficiently large η , the frontier $F_{\tau^*(\eta)}$ collapses to the distrust point: by Proposition 3, the designer chooses uninformative covariates. Below this cutoff, $F_{\tau^*(\eta)}$ is a nondegenerate polyline, with one segment for each group in which the designer relies on the AI. As η decreases, more groups enter the reliance set, adding segments to the polyline. At the same time, the designer relaxes the covariate bounds, so the corresponding segments become flatter. The stars trace the optimal partial reliance points $P^*(\eta)$ as η varies, and their upper-right envelope gives the full risk-reward frontier.

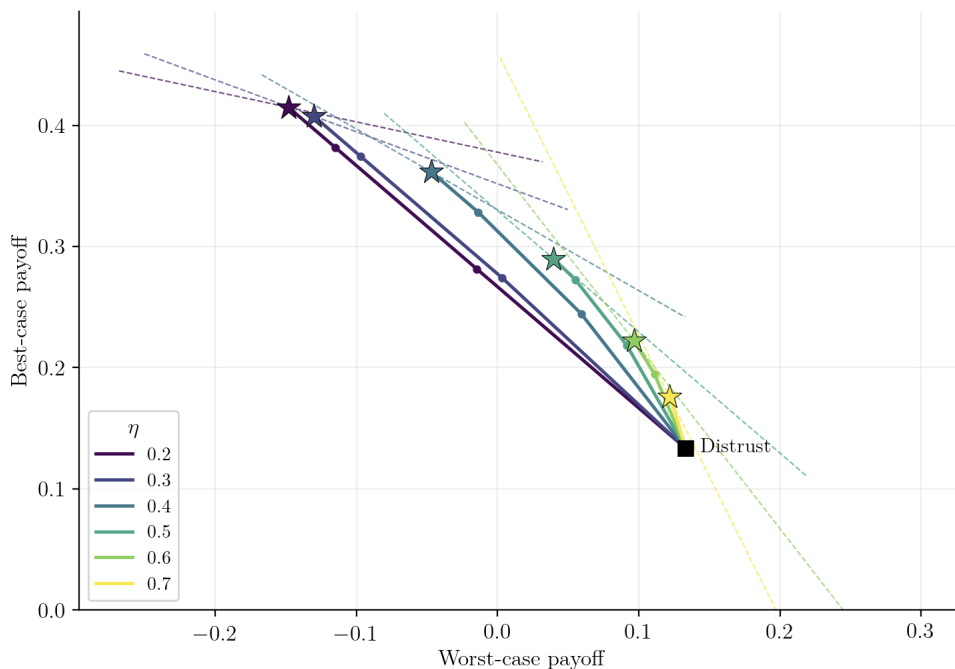


Figure 9: Each colored polyline is the fixed- η frontier $F_{\tau^*(\eta)}$, with kinks at partial-reliance points where one additional age group is delegated to the AI. As η decreases, the designer expands the reliance set and relaxes the optimal bounds within groups already in it, so the frontiers fan outward. Stars mark the optimal partial-reliance points $P^*(\eta)$, dashed lines show the supporting indifference lines, and the black curve connecting the stars is the full risk-reward frontier.

6 Extensions

6.1 Control of the covariate distribution

In the baseline model, Nature chooses the marginal distribution of X . Suppose now that, in addition to choosing τ , the designer also chooses for each group g an arbitrary marginal distribution $q_g \in \Delta(\mathcal{X})$ over a fixed non-singleton covariate set \mathcal{X} . The ambiguity set consists of all distributions $P \in \mathcal{P}(I)$ that satisfy the previous constraints and additionally $P(X = x | G = g) = q_g(x)$ for all x and g .

Proposition 4. *Fix τ . Allowing the designer to choose the marginal distribution of X does not change the risk-reward frontier: the resulting frontier is again F_τ .*

Thus choice of the marginal distribution of X adds no power beyond the choice of success-probability bounds and treatment rates. The proof, which is in the Online Appendix, shows that for each η , the designer optimally chooses the marginal distribution on X that a benevolent Nature would have chosen at the reliance point.

6.2 AI reports a distribution

Suppose that instead of asking the AI to choose an allocation rule a subject to a treatment rate, the designer asks it to report a joint distribution of (G, X, Y) in the ambiguity set $\mathcal{P}(I)$ and then treats those (g, x) for which the reported posterior satisfies $P(Y = 1 | G = g, X = x) \geq 1/2$ (thus removing the treatment-rate constraint). This makes the designer weakly worse off than in our baseline model.

To see this, observe that any treatment allocation induced by an admissible report can be replicated in our model by choosing, for each group g , the treatment rate equal to the mass treated under that report, and then allowing the AI to choose the corresponding treated set directly. So the designer in our model can match any payoffs achieved in this alternative model. But our model removes one source of downside risk that is present in the other: once the designer fixes r_g , a misaligned AI can no longer manipulate how many individuals are treated within group g , only *which* individuals are treated. Thus every payoff pair attainable in the alternative model is weakly dominated by a payoff pair attainable here.

6.3 Asymmetric treatment payoffs

If $u(1, 1) = \beta > 0$, $u(1, 0) = -\alpha < 0$, and $u(0, y) = 0$, the treatment threshold generalizes from $1/2$ to $\bar{p} := \frac{\alpha}{\alpha + \beta}$. Our main results continue to hold: Each group frontier is still a line segment from

D_g to R_g , the fixed- τ frontiers remain piecewise linear, optimal conditional-success probability bounds are asymmetric, and the extensive margin of reliance in each group is governed by a cutoff in η . However, the ordering of groups changes as follows. When $p_g < \min\{1/2, \bar{p}\}$, then groups are ordered by $\frac{\bar{p}-p_g}{1-p_g}$, and when $p_g \geq \max\{1/2, \bar{p}\}$ then they are ordered by $\frac{p_g-\bar{p}}{p_g}$.¹⁰ Intuitively, this ordering reflects a normalization of the distance from the treatment threshold, where the raw distance is divided by the amount of feasible movement in the relevant direction. Below \bar{p} , the default is not to treat, and so the relevant direction is towards 1 (reversing the default action); above \bar{p} , the default is to treat, so the relevant direction is towards 0. When $\alpha = \beta$, these normalized distances are the same increasing transformation of $|p_g - \frac{1}{2}|$ on both sides of the threshold, so they induce the ordering given in Theorem 2.

7 Conclusion

Our results have three implications for the governance of AI systems whose objective is not fully known. First, Proposition 1 cautions against relying on treatment-rate limits alone. When the AI’s inputs are unrestricted, the risk-reward frontier is a line segment with fixed slope. The designer can move along this frontier by adjusting treatment rates, but cannot improve the underlying tradeoff: each unit of best-case gain requires one unit of worst-case loss. Treatment-rate controls therefore limit exposure to AI influence, but they do not make delegation safer per unit of upside.

Second, Proposition 3 shows that input restrictions can improve the tradeoff between best-case and worst-case payoffs. By giving the AI access only to covariates whose informativeness is known to be bounded, the designer changes both what an aligned AI can achieve and what a misaligned AI can exploit. Proposition 3 characterizes the resulting restriction. For groups that are ex ante unlikely to benefit from treatment, the designer should allow information that can identify clear need for treatment while limiting information that can identify clear harm. For groups that are ex ante likely to benefit, the logic reverses. The lesson is not simply to make the AI less informed, but to control what the AI can be informative about.

Third, Theorem 2 identifies when reliance on the AI is most valuable. Reliance is least costly in groups whose baseline success probabilities are closest to the treatment threshold. In such groups, giving the AI discretion has relatively more upside than downside. This rule is moreover simple to implement, since it requires knowledge only of group-level baseline rates.

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

¹⁰Groups with p_g strictly between $1/2$ and \bar{p} require a separate analysis of the optimal bounds, which we omit.

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, 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 couple model uncertainty and alignment uncertainty by assuming Nature and the AI always act in concert. Future work could decouple these two sources of ambiguity and characterize frontiers under each separately, e.g. by applying α -MEU (Ghirardato et al. (2004)). Fourth, our $(\underline{\tau}, \bar{\tau})$ are abstract and unrestricted choice variables. One could derive a set of feasible $(\underline{\tau}, \bar{\tau})$ from the statistical properties of a chosen subset of covariates (e.g., the size of their supports).

Appendix

A Proof of Lemma 1

Fix τ and let $I = (\mathcal{X}, \tau)$ for any $|\mathcal{X}| \geq 2$. Moreover define $I' = (\mathcal{X}', \tau)$ 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 $\pi_g(A = 1) = r_g$, and consistency with the designer’s ambiguity set requires $\pi_g(Y = 1) = p_g$ together with $\pi_g(Y = 1 \mid A = a) \in [\underline{\tau}_g, \bar{\tau}_g]$ for each $a \in \{0, 1\}$, since $A \perp\!\!\!\perp Y \mid (G, X)$ and $P^*(Y = 1 \mid G = g, X = x) \in [\underline{\tau}_g, \bar{\tau}_g]$ for every $x \in \mathcal{X}$.

We now argue those conditions are sufficient for feasibility of π_g . That is, any π_g satisfying these three properties 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 $P'(X = x^1 \mid G = g) = r_g$ and $P'(X = x^0 \mid G = g) = 1 - r_g$, 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.¹¹

B Proof of Proposition 1

This is the special case of Theorem 1 with $\tau = \tau_0 = (0, 1)$ for every group. Under these bounds, $r_g^* = p_g$ and $\underline{\tau}_g = 0$, so the group reliance point reduces to $R_g(\tau_0) = (w(p_g), b(p_g))$ 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{DR} .

C Proof of Theorem 1

Section C.1 characterizes the best- and worst-targeting payoffs $b_g(p, r)$ and $w_g(p, r)$ defined in Section 4.2. It also shows that the feasible set of worst-case and best-case payoff pairs in group g is $\text{co}\{(w_g(p, r), b_g(p, r)) : r \in [0, 1]\}$. Section C.1.1 proves Lemma 3, thus bounding this feasible set below the line segment $\overline{R_g D_g}$ as depicted in Figure 7. Section C.2 constructs explicit information environments and policies that attain every point on $\overline{R_g D_g}$, thus proving Proposition 2, and Section C.3 aggregates the group-level frontiers.

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 $(\underline{\tau}_g, \bar{\tau}_g) \in [0, p_g] \times [p_g, 1]$ satisfying $\underline{\tau}_g < \bar{\tau}_g$. Then*

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

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

¹¹Since the construction assigns $P'(X = x_1 | G = g) = r_g$ and $P'(X = x_0 | G = g) = 1 - r_g$, full support requires $r_g \in (0, 1)$. For $r_g \in \{0, 1\}$, use any strictly positive distribution over \mathcal{X}' with the constant allocation rule $a(g, x) \equiv r_g$, as in the boundary cases of Lemma C.2.

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

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

Let $\pi_1 := \pi(Y = 1 | A = 1)$ and $\pi_0 := \pi(Y = 1 | A = 0)$. 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

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

Second, τ -admissibility of π requires

$$\underline{\tau} \leq \pi_1 \leq \bar{\tau} \quad \text{and} \quad \underline{\tau} \leq \pi_0 \leq \bar{\tau}. \quad (\text{C.2})$$

First suppose $r > 0$. Then by (C.1), $\pi_1 = \frac{p - (1-r)\pi_0}{r}$. As π_0 ranges over $[\underline{\tau}, \bar{\tau}]$, the corresponding values of π_1 lie in the interval $\left[\frac{p - (1-r)\bar{\tau}}{r}, \frac{p - (1-r)\underline{\tau}}{r}\right]$. The feasible set for π_1 is then the closed interval $\left[\frac{p - (1-r)\bar{\tau}}{r}, \frac{p - (1-r)\underline{\tau}}{r}\right] \cap [\underline{\tau}, \bar{\tau}]$. Since the objective $r(2\pi_1 - 1)$ is monotone in π_1 , its extrema occur at the endpoints $\bar{\pi}_1 = \min\left\{\bar{\tau}, \frac{p - (1-r)\underline{\tau}}{r}\right\}$ and $\underline{\pi}_1 = \max\left\{\underline{\tau}, \frac{p - (1-r)\bar{\tau}}{r}\right\}$. Comparing the two arguments of $\bar{\pi}_1$, we have $\frac{p - (1-r)\underline{\tau}}{r} \geq \bar{\tau} \iff p \geq \underline{\tau} + r(\bar{\tau} - \underline{\tau}) \iff r \leq r^* := \frac{p - \underline{\tau}}{\bar{\tau} - \underline{\tau}}$. Thus

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

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

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

Therefore,

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

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

The following lemma shows that these best-case and worst-case targeting payoffs define the

extremal points of the feasible set of (\underline{v}, \bar{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 $\bar{\tau} := \bar{\tau}_g$. Also let $\underline{v}_{I,g}(\mathbf{r}) = \inf_{P^* \in \mathcal{P}(I)} \inf_{a \in \mathcal{A}(I, \mathbf{r}, P^*)} \mathbb{E}_{(P^*, a)}[u(A, Y) \mid G = g]$ and $\bar{v}_{I,g}(\mathbf{r}) = \sup_{P^* \in \mathcal{P}(I)} \sup_{a \in \mathcal{A}(I, \mathbf{r}, P^*)} \mathbb{E}_{(P^*, a)}[u(A, Y) \mid G = g]$ be the worst- and best-case payoffs conditional on group g . Then the set of attainable payoff pairs $(\underline{v}_{I,g}(\mathbf{r}), \bar{v}_{I,g}(\mathbf{r}))$ 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 $\pi(Y = 1 \mid A = a) \in [\underline{\tau}, \bar{\tau}]$ for each a , whenever this conditional probability is defined, since $A \perp\!\!\!\perp Y \mid (G, X)$ and $P^*(Y = 1 \mid G = g, X = x) \in [\underline{\tau}, \bar{\tau}]$ 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 $P^*(Y = 1 \mid G = g, X = x^1) = \pi(Y = 1 \mid A = 1)$ and $P^*(Y = 1 \mid G = g, X = x^0) = \pi(Y = 1 \mid A = 0)$, 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, $\bar{v}_{I,g}(r) = \sup_{\pi \in \Pi_g(p, r)} \mathbb{E}_\pi[u(A, Y)] = b_g(p, r)$.

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 $\pi(Y = 1 \mid A = 0) = p \in [\underline{\tau}, \bar{\tau}]$, 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 $P^*(Y = 1 \mid G = g, X = x) = p$ for all $x \in \mathcal{X}$, 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 $\underline{v}_{I,g}(0) = \bar{v}_{I,g}(0) = 0 = w_g(p, 0) = b_g(p, 0)$.

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 $\pi(Y = 1 \mid A = 1) = p \in [\underline{\tau}, \bar{\tau}]$, which again holds by construction. This law is feasible by choosing any strictly positive distribution of X conditional on $G = g$, setting $P^*(Y = 1 \mid G = g, X = x) = p$ for all $x \in \mathcal{X}$, 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 $\underline{v}_{I,g}(1) = \bar{v}_{I,g}(1) = 2p - 1 = w_g(p, 1) = b_g(p, 1)$. Combining the interior and boundary cases proves the claim. \square

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 $\bar{v}_g \leq b_g(p, r^*)$.

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

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

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

In all of these cases, $c_1 = b - d = 0$, $c_2 = d - w \geq 0$, and $c_3 = (b - w)d = (d - w)d$. So the diagonal inequality reduces to $(d - w)\bar{v} \leq (d - w)d$. If $d - w = 0$, this is trivial. If $d - w > 0$, it is equivalent to $\bar{v} \leq d$, which follows from part (a) because $b = d$ (by (C.3)).

It remains to consider $\underline{\tau} < \frac{1}{2} < \bar{\tau}$ and $r^* \in (0, 1)$. For each fixed rate $r \in [0, 1]$, Lemma C.2 implies that the attainable conditional payoff pair is $\Gamma(r) := (W(r), B(r)) := (w_g(p, r), b_g(p, r))$. 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 $D := (d, d)$ and the reliance point $R := (w, b) = \Gamma(r^*)$.

Define $H(\underline{v}, \bar{v}) := (b - d)\underline{v} + (d - w)\bar{v} - (b - w)d$. Then $H = 0$ is exactly the line through D and R. Thus it is enough to show that $H(\Gamma(r)) \leq 0$ 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 $\Gamma(r^*) = R$, so $H(\Gamma(r^*)) = 0$. At $r = 0$, we have $\Gamma(0) = (0, 0)$, so $H(\Gamma(0)) = -(b - w)d \leq 0$, because $b \geq w$ and $d \geq 0$. At $r = 1$, we have $\Gamma(1) = (2p - 1, 2p - 1)$, so

$H(\Gamma(1)) = (b - w)(2p - 1 - d) \leq 0$, since $d = \max\{0, 2p - 1\}$.

It remains to check $r = 1 - r^*$. If $r^* \leq \frac{1}{2}$, 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 $\Gamma(r) = R + (r - r^*)(2\underline{\tau} - 1)(1, 1)$ for $r \in [r^*, 1 - r^*]$. Since $H(R) = 0$, 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 $2\bar{\tau} - 1$. Hence $\Gamma(r) = R + (r - r^*)(2\bar{\tau} - 1)(1, 1)$ for $r \in [1 - r^*, r^*]$. Again using $H(R) = 0$, we obtain $H(\Gamma(r)) = (r - r^*)(2\bar{\tau} - 1)(b - w) \leq 0$, because now $r - r^* \leq 0$, $2\bar{\tau} - 1 > 0$, 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 $\bar{\tau} \geq P(Y = 1 \mid A = 1, G = g) \geq \underline{\tau}$. Suppose either $\bar{\tau} \leq 1/2$ or $\underline{\tau} \geq 1/2$. Then the designer cannot improve upon the distrust point D_g for any preference parameter η , since if $\bar{\tau} \leq 1/2$ then $P(Y = 1 \mid A = 1, G = g) \leq \bar{\tau} \leq 1/2$, implying $\mathbb{E}[u(A, Y) \mid G = g] \leq 0 = d$ (where this final equality follows because $\bar{\tau} \leq 1/2$ 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 $\bar{\tau} > 1/2 > \underline{\tau}$. Lemma 3 confines the feasible set to the intersection of three halfspaces in the (\underline{v}, \bar{v}) plane: $\{\bar{v} \leq b\}$, $\{\underline{v} \leq d\}$, and $\{c_1\underline{v} + c_2\bar{v} \leq c_3\}$. In particular, every feasible payoff pair is dominated by some point on the line segment \overline{RD} . See the shaded region of Figure 7.

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

It remains to show that the two endpoints are implementable. To implement the distrust point $D = (d, d)$, 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 $R = (w, b)$, choose the rate $r = r^* := \frac{p - \underline{\tau}_g}{\bar{\tau}_g - \underline{\tau}_g}$. By definition of r^* , there is an admissible two-point posterior distribution placing mass r^* on $\bar{\tau}_g$ and mass $1 - r^*$

on $\underline{\tau}_g$. An aligned AI treats the $\bar{\tau}_g$ -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 .

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

C.3 Constructing the Full Frontier

For each group $g \in \mathcal{G}$, let $R_g = (w_g(p_g, r_g^*), b_g(p_g, r_g^*))$ and $D_g = (d_g, d_g)$ 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, \mathbf{r}) and any (P^*, a) where $P^* \in \mathcal{P}(I)$ and a is admissible. For each group g , let $z_g = (\underline{v}_g, \bar{v}_g) \in C_g$ 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 $\underline{v}_I(\mathbf{r}) = \sum_{g \in \mathcal{G}} \mu_g \underline{v}_g$ and $\bar{v}_I(\mathbf{r}) = \sum_{g \in \mathcal{G}} \mu_g \bar{v}_g$. Hence $(\underline{v}_I(\mathbf{r}), \bar{v}_I(\mathbf{r})) \in \sum_{g \in \mathcal{G}} \mu_g C_g$.

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 implements 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

$$C_\tau = \sum_{g \in \mathcal{G}} \mu_g C_g := \left\{ \sum_{g \in \mathcal{G}} \mu_g z_g : z_g \in C_g \right\}. \quad (\text{C.4})$$

We now characterize the Pareto frontier of C_τ . Fix $\eta \in [0, 1]$ and consider the supporting functional $\Lambda_\eta(\underline{v}, \bar{v}) := \eta \underline{v} + (1 - \eta) \bar{v}$. Because C_τ is the weighted Minkowski sum (C.4) and Λ_η is linear, maximization separates across groups:

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

As shown in Section C.2, the frontier for group g is the line segment $F_g := \overline{R_g D_g}$. 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

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

Whenever $R_g \neq D_g$, this is equivalent to

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

Order groups as g_1, \dots, g_n so that $\Delta(g_1) \leq \Delta(g_2) \leq \dots \leq \Delta(g_n)$. For each $k = 0, 1, \dots, n$, let $G_k := \{g_1, \dots, g_k\}$, with the convention $G_0 = \emptyset$, and define

$$P^{(k)} := \sum_{g \in G_k} \mu_g R_g + \sum_{g \notin G_k} \mu_g D_g.$$

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

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 $k(\eta) \in \{0, \dots, n\}$ such that the maximizer of Λ_η is the point $P^{(k(\eta))}$ 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, \dots, g_n . Therefore the supported points of C_τ trace exactly the chain $\overline{DP^{(1)}} \cup \overline{P^{(1)}P^{(2)}} \cup \dots \cup \overline{P^{(n-1)}R}$. Since C_τ is compact and convex, every Pareto-undominated boundary point is supported by some linear functional Λ_η . Hence the risk-reward frontier is exactly $F_\tau = \overline{DP^{(1)}} \cup \overline{P^{(1)}P^{(2)}} \cup \dots \cup \overline{P^{(n-1)}R}$, 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 $\tau = (\underline{\tau}, \bar{\tau}) := (\underline{\tau}_g, \bar{\tau}_g)$ for this group is:*

Case 1: $p < 1/2$. Define $\underline{\eta}(p) := \frac{1-p}{2} < \frac{1}{2(1-p)} =: \bar{\eta}(p)$ and let

$$\underline{\tau}^\circ(\eta) = 1 - \sqrt{\frac{1-p}{2\eta}} \quad \forall \eta \in [\underline{\eta}(p), \bar{\eta}(p)].$$

Then an optimal choice of τ is

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

and this solution is unique for $\eta < \bar{\eta}(p)$.

Case 2: $p > 1/2$. Define $\underline{\eta}(p) := \frac{p}{2} < \frac{1}{2p} =: \bar{\eta}(p)$ and let

$$\bar{\tau}^\circ = \sqrt{\frac{p}{2\eta}} \quad \forall \eta \in [\underline{\eta}(p), \bar{\eta}(p)].$$

Then an optimal choice of τ is

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

and this solution is unique for $\eta < \bar{\eta}(p)$.

Case 3: $p = 1/2$. Both solutions in cases 1 and 2 are optimal.

Proof. Because τ_g is chosen independently across groups and the designer's objective $\eta\underline{v} + (1 - \eta)\bar{v}$ 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 $D = ((2p - 1)_+, (2p - 1)_+)$, where $(x)_+ := \max\{0, x\}$. and the reliance point $R_g(\tau) = (w(\tau), b(\tau))$ where

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

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

We prove the $p < 1/2$ case. Here the distrust point $D = (0, 0)$ yields a payoff of zero, which the designer can implement by setting either $\underline{\tau} = p$ or $p = \bar{\tau}$ (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

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

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

We prove in Online Appendix OA.2 that $r^* \leq \frac{1}{2}$ (equivalently, $p \leq (\underline{\tau} + \bar{\tau})/2$). Thus $w(\tau) = r^*(2\underline{\tau} - 1)$ and $b(\tau) = r^*(2\bar{\tau} - 1)$. Using $p = (1 - r^*)\underline{\tau} + r^*\bar{\tau}$, we can rewrite

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

hence

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

For fixed $\tau < 1/2$, this expression is strictly increasing in $\bar{\tau}$; thus any nondegenerate reliance optimum sets $\bar{\tau} = 1$. The problem then reduces to maximizing

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

over $\tau \in [0, p]$. This objective is strictly concave in τ with derivative

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

Hence the unique maximizer is

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

The corresponding payoff is

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

The argument for $p > 1/2$ is symmetric, with the roles of τ and $\bar{\tau}$ exchanged. With $p > 1/2$, the designer optimally sets $\tau^* = 0$ throughout and

$$\bar{\tau}^* = \begin{cases} 1, & \eta \leq \frac{p}{2}, \\ \sqrt{\frac{p}{2\eta}}, & \frac{p}{2} < \eta < \frac{1}{2p}, \\ p, & \eta \geq \frac{1}{2p}. \end{cases}$$

E Proof of Theorem 2

The proof characterizes the slope of each group g frontier when τ_g is optimally chosen, and uses this to show that ordering groups by this slope is equivalent to ordering them by $|p_g - 1/2|$. It then shows that each designer's reliance set consists of the first groups in this ordering, where the number of groups 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 $\delta_g = |p_g - 1/2|$, $\underline{\eta}_g = \frac{1+2\delta_g}{4}$, and $\bar{\eta}_g = \frac{1}{1+2\delta_g}$. Then the slope of the group frontier line segment $\overline{R_g D_g}$ is

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

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

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

Proof. Suppose $p \leq 1/2$. By Proposition D.1, there exist $0 \leq \underline{\eta}_g < \bar{\eta}_g \leq 1$ such that for $\eta < \underline{\eta}_g$ the designer optimally chooses $(\underline{\tau}^*(\eta), \bar{\tau}^*(\eta)) = (0, 1)$, so $R_g = (-p, p)$ and the distrust point is $D_g = (0, 0)$, with slope -1 , yielding Part (a). When $\underline{\eta}_g \leq \eta < \bar{\eta}_g$ then the designer optimally chooses $(\underline{\tau}(\eta), \bar{\tau}(\eta)) = \left(1 - \sqrt{\frac{1-p}{2\eta}}, 1\right)$ so the reliance point is $R_g = \left(\left(\frac{p-\underline{\tau}^\circ}{1-\underline{\tau}^\circ}\right)(2\underline{\tau}^\circ - 1), \frac{p-\underline{\tau}^\circ}{1-\underline{\tau}^\circ}\right)$ while the distrust point is $D = (0, 0)$. (Recall from Proposition D.1 that $\underline{\tau}^\circ = 1 - \sqrt{\frac{1-p}{2\eta}}$.) So the slope is

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

using $p \leq \frac{1}{2}$ in the final equality. This yields Part (b). When $\eta \geq \bar{\eta}_g$ then the designer optimally chooses $(\underline{\tau}(\eta), \bar{\tau}(\eta)) = (p, 1)$ so $R_g = D_g = (0, 0)$, yielding Part (c).

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

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

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

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

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

We now use Corollary E.1 to complete the proof. First we show that there exists a cutoff index $T(\eta)$ such that the reliance set of the designer with preference parameter η is simply the first $T(\eta)$

groups in the ordering by $\delta_g = |p_g - 1/2|$. To do this, we show that

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

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 $\eta \geq \bar{\eta}_g$, then it must be that $\eta \geq \bar{\eta}_{g'}$ because $\bar{\eta}_g$ is decreasing in δ_g . So $\Delta_g = 0 = \Delta_{g'}$ and we are done. If $\eta < \bar{\eta}_g$ and $\eta \geq \bar{\eta}_{g'}$ then $\Delta(g', \eta) = 0 \geq \Delta(g, \eta)$.

It remains to consider $\underline{\eta}_g \leq \eta < \bar{\eta}_g$ and $\eta < \bar{\eta}_{g'}$. When $\underline{\eta}_{g'} \leq \eta < \bar{\eta}_{g'}$, 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 $g^{(1)}, \dots, g^{(|\mathcal{G}|)}$ such that $\delta_{g^{(1)}} \leq \delta_{g^{(2)}} \leq \dots \leq \delta_{g^{(|\mathcal{G}|)}}$. By (E.1), the corresponding slopes satisfy $\Delta(g^{(1)}, \eta) \leq \Delta(g^{(2)}, \eta) \leq \dots \leq \Delta(g^{(|\mathcal{G}|)}, \eta)$. Recall that

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

is the condition that the optimal point is $R_g(\eta)$ rather than D_g . Thus there exists $T(\eta) \in \{0, 1, \dots, |\mathcal{G}|\}$ such that the set of indices for which (E.2) holds are those satisfying $j \leq T(\eta)$. Equivalently, the optimizer selects $R_{g^{(j)}}(\eta)$ for $j \leq T(\eta)$ and selects $D_{g^{(j)}}$ for $j > T(\eta)$. 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 $T(\eta)$ is monotone in η . We will show that (E.2) holds exactly when $\eta < \bar{\eta}_g$ (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 $\underline{\eta}_g \leq \eta < \bar{\eta}_g$, then by the corollary

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

and a direct rearrangement shows that

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

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

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

F Proof of Theorem 3

Let $K := \overline{\text{co}}(\bigcup_{\tau} C(\tau))$, where $C(\tau)$ is the feasible set given fixed success-probability bounds τ . The risk-reward frontier is the Pareto frontier of K . We show that this frontier is also the Pareto frontier of the closed convex hull of $S := \{\mathbf{P}^*(\eta) : \eta \in (0, 1)\}$.

For each $\eta \in (0, 1)$, define the linear functional $\Lambda_{\eta}(\underline{v}, \bar{v}) := \eta \underline{v} + (1 - \eta) \bar{v}$. Fix any η . Proposition 3 determines an optimal choice $\tau^*(\eta)$, and Theorem 1 applied to $\tau^*(\eta)$ shows that the optimal partial-reliance point $\mathbf{P}^*(\eta)$ maximizes Λ_{η} over the feasible set $C(\tau^*(\eta))$. Since $\tau^*(\eta)$ was chosen to maximize $\Lambda_{\eta}(\underline{v}, \bar{v})$ over all τ , it follows that $\mathbf{P}^*(\eta)$ also maximizes Λ_{η} over the full feasible set and hence over K . Therefore $\max_{z \in K} \Lambda_{\eta}(z) = \Lambda_{\eta}(\mathbf{P}^*(\eta))$.

Since $S \subseteq K$ and $\mathbf{P}^*(\eta) \in S$, the point $\mathbf{P}^*(\eta)$ also attains the maximum of Λ_{η} over S . Thus

$$\max_{z \in S} \Lambda_{\eta}(z) = \Lambda_{\eta}(\mathbf{P}^*(\eta)) = \max_{z \in K} \Lambda_{\eta}(z).$$

A linear functional has the same maximum on a set as on its closed convex hull, so for every $\eta \in (0, 1)$, $\max_{z \in K} \Lambda_{\eta}(z) = \max_{z \in \overline{\text{co}}(S)} \Lambda_{\eta}(z)$. Thus K and $\overline{\text{co}}(S)$ have the same supporting lines in every nonnegative direction.

Finally, the Pareto frontier of any closed convex subset of \mathbb{R}^2 is exactly the union of the faces supported by nonnegative linear functionals. Hence K and $\overline{\text{co}}(S)$ have the same Pareto frontier. This proves that the risk-reward frontier is the set of Pareto-undominated points in $\overline{\text{co}}(\{\mathbf{P}^*(\eta) : \eta \in (0, 1)\})$, as claimed.

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.
- AUTOR, D., A. CAPLIN, D. J. MARTIN, AND P. MARX (2025): “Misaligned by Design: Incentive Failures in Machine Learning,” Tech. rep., National Bureau of Economic Research.
- 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*.
- BERGEMANN, D. AND S. MORRIS (2005): “Robust Mechanism Design,” *Econometrica*, 73, 1771–1813.
- BOMMASANI, R., D. A. HUDSON, E. ADELI, R. ALTMAN, S. ARORA, S. VON ARX, M. S. BERNSTEIN, J. BOHG, A. BOSSELUT, E. BRUNSKILL, ET AL. (2021): “On the Opportunities and Risks of Foundation Models,” *arXiv preprint arXiv:2108.07258*.

- BROOKS, B. AND S. DU (2021): “Optimal Auction Design With Common Values: An Informationally Robust Approach,” *Econometrica*, 89, 1313–1360.
- BURRELL, J. (2016): “How the Machine “Thinks”: Understanding Opacity in Machine Learning Algorithms,” *Big Data & Society*, 3.
- CARLSMITH, J. (2023): “Scheming AIs: Will AIs Fake Alignment During Training in Order to Get Power?” *arXiv preprint arXiv:2311.08379*.
- CARROLL, G. (2017): “Robustness and Separation in Multidimensional Screening,” *Econometrica*, 85, 453–488.
- CHEN, E. O., A. GHERSENGORIN, AND S. PETERSEN (2024): “Imperfect Recall and AI Delegation,” Working Paper 30-2024, Global Priorities Institute, University of Oxford.
- CHRISTIANO, P. (2019): “What Failure Looks Like,” AI Alignment Forum.
- CHRISTIANO, P., B. SHLEGERIS, AND D. AMODEI (2016): “Concrete Problems in AI Safety,” *arXiv preprint arXiv:1606.06565*.
- COLLINA, N., S. GOEL, A. ROTH, E. RYU, AND M. SHI (2025): “Emergent Alignment via Competition,” *arXiv preprint arXiv:2509.15090* (revised February 2026).
- CROSS, P. J. AND C. F. MANSKI (2002): “Regressions, Short and Long,” *Econometrica*, 70, 357–368.
- DENISON, C., M. MACDIARMID, F. BAREZ, D. DUVENAUD, S. KRAVEC, S. MARKS, N. SCHIEFER, R. SOKLASKI, A. TAMKIN, ET AL. (2024): “Sycophancy to Subterfuge: Investigating Reward-Tampering in Large Language Models,” *arXiv preprint arXiv:2406.10162*.
- DESSEIN, W. (2002): “Authority and Communication in Organizations,” *Review of Economic Studies*, 69, 811–838.
- DOSHI-VELEZ, F. AND B. KIM (2017): “Towards a Rigorous Science of Interpretable Machine Learning,” *arXiv preprint arXiv:1702.08608*.
- 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.
- EQUAL EMPLOYMENT OPPORTUNITY COMMISSION, CIVIL SERVICE COMMISSION, DEPARTMENT OF LABOR, AND DEPARTMENT OF JUSTICE (1978): “Uniform Guidelines on Employee Selection Procedures,” 43 Federal Register 38290.
- FEDERAL FINANCIAL INSTITUTIONS EXAMINATION COUNCIL (2009): “Interagency Fair Lending Examination Procedures,” <https://www.ffiec.gov/PDF/fairlend.pdf>.
- FRANKEL, A. (2021): “Selecting applicants,” *Econometrica*, 89, 615–645.
- GHIRARDATO, P., F. MACCHERONI, AND M. MARINACCI (2004): “Differentiating Ambiguity and Ambiguity Attitude,” *Journal of Economic Theory*, 118, 133–173.
- 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*.
- HOLMSTRÖM, B. (1984): “On the theory of delegation,” *Discussion Paper*.
- HUBINGER, E., C. DENISON, J. MU, M. LAMBERT, M. TONG, M. MACDIARMID, T. LANHAM, D. M. ZIEGLER, ET AL. (2024): “Sleepers Agents: Training Deceptive LLMs that Persist Through Safety Training,” *arXiv preprint arXiv:2401.05566*.
- HUBINGER, E., C. VAN MERWIJK, V. MIKULIK, J. SKALSE, AND S. GARRABRANT (2019): “Risks from Learned Optimization in Advanced Machine Learning Systems,” *arXiv preprint arXiv:1906.01820*.
- 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?” in *The Economics of Transformative AI*, ed. by A. K. Agrawal, E. Brynjolfsson, and A. Korinek, University of Chicago Press, chap. 16.
- 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*, 134, 1401–1467.
- LIN, X. AND C. LIU (2024): “Credible persuasion,” *Journal of Political Economy*, 132, 2228–2273.
- LIPTON, Z. C. (2018): “The Mythos of Model Interpretability,” *Queue*, 16, 31–57.
- 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.
- METR (2026): “Frontier Risk Report (February to March 2026),” <https://metr.org/blog/2026-05-19-frontier-risk-report/>.
- MONTIEL OLEA, J. L., C. QIU, AND J. STOYE (2026): “Decision Theory for Treatment Choice Problems with Partial Identification,” *The Review of Economic Studies*, advance access: doi:10.1093/restud/rdag015.
- NGO, R., L. CHAN, AND S. MINDERMANN (2025): “The alignment problem from a deep learning perspective,” *arXiv preprint arXiv:2209.00626v8*.
- OBERMEYER, Z., B. POWERS, C. VOGELI, AND S. MULLAINATHAN (2019): “Dissecting Racial Bias in an Algorithm Used to Manage the Health of Populations,” *Science*, 366, 447–453.
- 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.

Online Appendix

OA.1 Proof of Proposition 4

Fix τ . Consider a fixed group g and write $p := p_g$, $\underline{\tau} := \underline{\tau}_g$, and $\bar{\tau} := \bar{\tau}_g$. Let $q = (q(x))_{x \in \mathcal{X}}$ be a designer-chosen marginal distribution for X in this group, and let $a : \mathcal{X} \rightarrow [0, 1]$ be an allocation rule satisfying $\sum_{x \in \mathcal{X}} q(x)a(x) = r$. For this fixed q and a , Nature chooses conditional success probabilities $m(x) := P(Y = 1 \mid G = g, X = x)$ subject to $m(x) \in [\underline{\tau}, \bar{\tau}] \forall x \in \mathcal{X}$, and $\sum_{x \in \mathcal{X}} q(x)m(x) = p$. Define

$$B_q(a) := \sup_m \sum_{x \in \mathcal{X}} q(x)a(x)(2m(x) - 1)$$

and

$$W_q(a) := \inf_m \sum_{x \in \mathcal{X}} q(x)a(x)(2m(x) - 1),$$

where the supremum and infimum are over all such posterior vectors m . For this fixed marginal q and rate r , write $\bar{v}_q(r) := \sup_a B_q(a)$ and $\underline{v}_q(r) := \inf_a W_q(a)$, where the supremum and infimum are over allocation rules a satisfying $\sum_{x \in \mathcal{X}} q(x)a(x) = r$.

We claim that for every admissible a ,

$$(W_q(a), B_q(a)) \in \text{co}\left\{(w_g(p, s), b_g(p, s)) : s \in [0, 1]\right\}. \quad (\text{OA.1.1})$$

To prove this, let $\rho := \frac{p - \underline{\tau}}{\bar{\tau} - \underline{\tau}} \in [0, 1]$. For the best case, write $m(x) = \underline{\tau} + (\bar{\tau} - \underline{\tau})z(x)$ with $z(x) \in [0, 1]$. Then the mean constraint becomes $\sum_x q(x)z(x) = \rho$, and maximizing $B_q(a)$ is equivalent to maximizing $\sum_x q(x)a(x)z(x)$ over $z(\cdot) \in [0, 1]^{\mathcal{X}}$ subject to the same mean constraint.

Since \mathcal{X} is finite and the feasible set is compact, a maximizer exists. Moreover, this maximizer must put z -mass on higher-allocation states before lower-allocation states. To see this, suppose towards contradiction that $a(x) > a(y)$ and, for a feasible z , we have $z(x) < 1$ and $z(y) > 0$ with $q(x), q(y) > 0$. Let

$$\varepsilon \leq \min\{q(x)(1 - z(x)), q(y)z(y)\}$$

be positive, and define \tilde{z} by increasing $z(x)$ by $\varepsilon/q(x)$, decreasing $z(y)$ by $\varepsilon/q(y)$, and leaving all other coordinates unchanged. Then $\tilde{z} \in [0, 1]^{\mathcal{X}}$ and

$$\sum_u q(u)\tilde{z}(u) = \sum_u q(u)z(u),$$

while the objective changes by

$$\sum_u q(u)a(u)\tilde{z}(u) - \sum_u q(u)a(u)z(u) = \varepsilon(a(x) - a(y)) > 0.$$

Thus z was not a maximizer. Coordinates with $q(x) = 0$ do not affect either the constraint or the objective and may be assigned arbitrarily. It follows that there is a cutoff c such that $z(x) = 1$ if

$a(x) > c$ and $z(x) = 0$ if $a(x) < c$, with values on $\{x : a(x) = c\}$ chosen to satisfy $\sum_x q(x)z(x) = \rho$.

Now define

$$E_t := \{x \in \mathcal{X} : a(x) \geq t\}$$

and $s(t) := \sum_{x \in E_t} q(x)$. Since $a(x) \in [0, 1]$,

$$a(x) = \int_0^1 \mathbb{1}\{a(x) \geq t\} dt.$$

The cutoff form implies the needed nested-set identity. If $t > c$, then E_t contains only states with $a(x) > c$, so $z(x) = 1$ on E_t and $s(t) \leq \rho$. If $t \leq c$, then E_t contains every state with $z(x) > 0$, so the qz -mass in E_t is the full mass ρ and $s(t) \geq \rho$. Therefore, for every $t \in [0, 1]$,

$$\sum_{x \in E_t} q(x)z(x) = \min\{s(t), \rho\}.$$

Consequently

$$\begin{aligned} B_q(a) &= \int_0^1 \sum_{x \in E_t} q(x)(2m(x) - 1) dt \\ &= \int_0^1 [s(t)(2\bar{\tau} - 1) + 2(\bar{\tau} - \underline{\tau}) \min\{s(t), \rho\}] dt \\ &= \int_0^1 b_g(p, s(t)) dt, \end{aligned}$$

where the last equality uses the formula for b_g in Lemma C.1.

For the worst case, again write $m(x) = \underline{\tau} + (\bar{\tau} - \underline{\tau})z(x)$ and set $y(x) := 1 - z(x)$. Then minimizing $W_q(a)$ is equivalent to maximizing $\sum_x q(x)a(x)y(x)$ subject to $y(x) \in [0, 1]$ and $\sum_x q(x)y(x) = 1 - \rho$. The corresponding cutoff property follows from the same mass shift. If $a(x) > a(x')$, $y(x) < 1$, $y(x') > 0$, and $q(x), q(x') > 0$, shift a positive amount

$$\varepsilon \leq \min\{q(x)(1 - y(x)), q(x')y(x')\}$$

of qy -mass from x' to x . This preserves $\sum_x q(x)y(x)$ and changes the objective by $\varepsilon\{a(x) - a(x')\} > 0$. Hence, on positive- q states, no maximizer can put positive y -mass on a lower-allocation state while a higher-allocation state has y below one. Therefore there is a cutoff d such that $y(x) = 1$ if $a(x) > d$, and $y(x) = 0$ if $a(x) < d$, with values on $\{x : a(x) = d\}$ chosen to satisfy $\sum_x q(x)y(x) = 1 - \rho$. As above, if $t > d$, then E_t contains only states with $y(x) = 1$ and $s(t) \leq 1 - \rho$; if $t \leq d$, then E_t contains all states with positive y -mass and the qy -mass in E_t is $1 - \rho$. Consequently, for every t ,

$$\sum_{x \in E_t} q(x)y(x) = \min\{s(t), 1 - \rho\}.$$

Hence

$$\sum_{x \in E_t} q(x)z(x) = s(t) - \min\{s(t), 1 - \rho\} = (s(t) + \rho - 1)_+,$$

and therefore

$$\begin{aligned}
W_q(a) &= \int_0^1 \sum_{x \in E_t} q(x)(2m(x) - 1) dt \\
&= \int_0^1 \left[s(t)(2\bar{\tau} - 1) + 2(\bar{\tau} - \underline{\tau})(s(t) + \rho - 1)_+ \right] dt \\
&= \int_0^1 w_g(p, s(t)) dt,
\end{aligned}$$

again by Lemma C.1. This proves (OA.1.1).

Now let a^+ be an allocation rule attaining the best-case payoff under the chosen q , so $\bar{v}_q(r) = B_q(a^+)$. The worst-case payoff under the same marginal q is

$$\underline{v}_q(r) = \inf_a W_q(a) \leq W_q(a^+).$$

By (OA.1.1), the point $(W_q(a^+), B_q(a^+))$ lies in the baseline convex hull

$$\text{co}\left\{ (w_g(p, s), b_g(p, s)) : s \in [0, 1] \right\}.$$

It has the same best-case coordinate as $(\underline{v}_q(r), \bar{v}_q(r))$ and a weakly larger worst-case coordinate. Hence every payoff pair generated by a designer-chosen marginal q is weakly dominated by a point in the baseline group feasible set. So designer control of q cannot create a new Pareto-undominated group payoff outside the baseline frontier.

Conversely, every point in the baseline group feasible set remains attainable when the designer chooses q . Fix any $s \in [0, 1]$. If $s \in (0, 1)$, choose distinct covariate values $x^1, x^0 \in \mathcal{X}$ and set $q(x^1) = s$, $q(x^0) = 1 - s$, and $q(x) = 0$ for all other x . Choose treatment rate $r = s$, and let the AI treat exactly those individuals with $X = x^1$. Then, exactly as in the reverse inclusion in Lemma C.2, any joint law $\pi \in \Pi_g(p, s)$ can be implemented by setting

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

Therefore the best- and worst-case payoffs under this designer-chosen marginal are $(w_g(p, s), b_g(p, s))$. If instead $s \in \{0, 1\}$, the same endpoint payoff pair is attained by any marginal q together with the corresponding constant action, exactly as in the boundary cases of Lemma C.2. Ex ante randomization then yields the entire convex hull

$$\text{co}\left\{ (w_g(p, s), b_g(p, s)) : s \in [0, 1] \right\}.$$

So the single-group feasible set with designer-chosen marginals coincides with the baseline one.

Finally, return to the full model with arbitrary \mathcal{G} . Because the choice of q_g and the treatment-rate constraint are both group-specific, the same decomposition as in (C.4) applies: the aggregate feasible set is the weighted Minkowski sum of the group feasible sets. Since those group feasible

sets are unchanged for every g , the aggregate feasible set is still C_τ . Its Pareto frontier is therefore still F_τ , proving the proposition.

OA.2 Completion of Proof of Proposition 3

Call the case $r^* \leq \frac{1}{2}$ analyzed above Regime (A). Consider now a second Regime (B): $r^* > \frac{1}{2}$ (equivalently $p > \frac{\underline{\tau} + \bar{\tau}}{2}$). We will show that the designer never chooses a $(\underline{\tau}, \bar{\tau})$ in this regime. By assumption, $\bar{\tau} > \frac{1}{2}$, in which case we have

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

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

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

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

With $\underline{\tau} = 0$, the condition $r^* > \frac{1}{2}$ becomes $p/\bar{\tau} > \frac{1}{2}$, i.e. $\bar{\tau} < 2p$. Hence Regime (B) reduces to choosing $\bar{\tau} \in [p, 2p]$ to maximize $U_R(0, \bar{\tau}) = \frac{p}{\bar{\tau}}(2\bar{\tau} - 1) - 2\eta(\bar{\tau} - p) = 2p(1 + \eta) - \left(\frac{p}{\bar{\tau}} + 2\eta\bar{\tau}\right)$. Because the sum of two non-negative numbers is at least twice their geometric mean, $\frac{p}{\bar{\tau}} + 2\eta\bar{\tau} \geq 2\sqrt{\frac{p}{\bar{\tau}} \cdot 2\eta\bar{\tau}} = 2\sqrt{2\eta p}$. Equality holds if and only if the two terms are equal, that is, when $\frac{p}{\bar{\tau}} = 2\eta\bar{\tau}$, or equivalently $\bar{\tau} = \sqrt{\frac{p}{2\eta}}$. Accounting for the constraint $\bar{\tau} \in [p, 2p]$ yields

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

where we use that $\bar{\tau} = \sqrt{p/(2\eta)}$ is feasible only when $p \leq \sqrt{\frac{p}{2\eta}} \leq 2p \iff \frac{1}{8p} \leq \eta \leq \frac{1}{2p}$. The last line corresponds to the boundary choice $\bar{\tau} = p$, which yields the distrust point.

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

Case 1: $\eta \leq \frac{1}{8p}$. Then $U_A^* - U_B^* \geq p(1 - 2\eta) - \left(2p(1 - \eta) - \frac{1}{2}\right) = \frac{1}{2} - p \geq 0$.

Case 2: $\frac{1}{8p} \leq \eta < \frac{1}{2(1-p)}$. Here $U_B^* = 2p(1 + \eta) - 2\sqrt{2\eta p}$. Moreover, for $p \leq \frac{1}{2}$, $\frac{1}{8p} \geq \frac{1-p}{2}$, 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 $U_A^* = \left(1 - \sqrt{2\eta(1-p)}\right)^2$. Therefore

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

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

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

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, $\frac{d}{d\eta} U_B^* = 2p - \sqrt{\frac{2p}{\eta}} \leq 0$ 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 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).