Algorithm Design and Physician Liability

Shujie Luan Shubhranshu Singh Tinglong Dai*

This version: August 12, 2026

Abstract

A single clinical algorithm can deliver unequal accuracy across patient groups, and concern about such disparity has grown as artificial intelligence (AI) spreads through clinical decision-making. In response, a liability rule introduced in the United States holds healthcare providers responsible when their reliance on disparate algorithms contributes to erroneous clinical decisions. We examine how such liability considerations reshape (i) an AI firm's algorithm design decisions that drive group-specific accuracy and (ii) a physician's decisions to use AI in healthcare delivery. The AI firm designs an algorithm for two patient groups, and improving accuracy for the disadvantaged group is more costly. The physician (who remains the accountable decision-maker) then decides whether to consult AI, weighing the reduction in clinical uncertainty against expected liability exposure when AI errors disproportionately affect the disadvantaged group. We find the liability rule can induce disparate *use* of AI: the physician may reduce AI use overall and, over an intermediate range of liability, rely on AI less for disadvantaged patients. The effect is non-monotone. As liability increases, the physician's use of AI for disadvantaged patients first declines, then rises as the firm reallocates investment toward reducing disparity or switches to an equal-accuracy design. Mandating equal algorithmic accuracy across patient groups can then inadvertently harm both groups, because a uniform accuracy requirement distorts the firm's investment incentives and the physician's equilibrium AI-use decisions.

Key words: clinical algorithms; product design; algorithmic disparity; liability; human–AI interaction

*Shujie Luan is affiliated with The University of Western Australia, UWA Business School, Perth, Australia; Shubhranshu Singh and Tinglong Dai are affiliated with Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202. The coauthors contributed equally to this work. Correspondence may be addressed to: shujie.luan@uwa.edu.au (SL); shubhranshu.singh@jhu.edu (SS); dai@jhu.edu (TD).

1. Introduction

Artificial intelligence (AI) is reshaping expert decision-making, yet most research treats the two sides of the AI lifecycle separately. One strand asks how firms design algorithms under technical and regulatory constraints (e.g., Diao et al. 2023, Israeli 2018, Iyer and Ke 2024); another asks how human experts adopt, trust, or strategically act on algorithmic advice (e.g., Dai and Singh 2025, Dietvorst et al. 2018, McLaughlin and Spiess 2022). In practice, these margins are jointly determined: the rules that govern how practitioners use AI feed back into how firms build it. This interdependence is especially consequential when regulation targets the point of use, whereas the performance disparities that trigger liability are shaped upstream during design. A feedback loop arises in which (i) a firm designs product quality for different customer segments and (ii) an expert intermediary, who faces policy constraints, decides whether the product is used.

Clinical decision support offers a paradigmatic setting in which this feedback loop is first-order. Although "clinical algorithms" once denoted transparent flow charts, advances in AI have pushed the frontier toward high-performing but often opaque systems (Gottlieb 2024, Green and Defoe 1978, Margolis 1983). As of early 2026, the U.S. Food and Drug Administration (FDA) has authorized over 1,300 AI-based medical devices (FDA 2025). These AI tools can improve screening, diagnosis, and treatment decisions (Leong et al. 2023, Rajpurkar et al. 2022, Topol 2019), yet their diffusion into routine care remains uneven (Abràmoff et al. 2024, Wu et al. 2023). A central barrier to adoption is the concern regarding *algorithmic disparity*—systematic differences in predictive performance across protected groups—which raises equity concerns and exposes clinicians and hospitals to significant legal risk (Mehrabi et al. 2021, Obermeyer et al. 2019, 2021). Reflecting these concerns, the Centers for Medicare & Medicaid Services (CMS) has extended Section 1557 nondiscrimination requirements to clinical decision support tools under § 92.210. This rule prohibits discriminatory use and explicitly requires covered entities to make reasonable efforts to identify and mitigate disparity, thereby placing the burden of algorithmic accountability directly on providers (CMS 2022, 2024). Because most healthcare providers in the U.S. context are physicians, we use the terms "physician" and "provider" interchangeably in the remainder of the paper.

Motivated by this policy shift, we study how deployment-facing nondiscrimination liability changes incentives in both the development and use of clinical AI when algorithms

perform unevenly across patient groups. To isolate this mechanism, we abstract from broader liability channels—most notably malpractice (Price et al. 2019)—and focus on the margin where this regulation is most directly enforced: clinicians' reliance on decision-support tools in patient care. Our model links upstream technology design to downstream clinical delivery in a two-stage structure. First, an AI firm chooses accuracy for two patient groups: an advantaged group for whom performance gains are relatively inexpensive, and a disadvantaged group for whom gains are costlier because of data sparsity and measurement frictions (Chen et al. 2018, Mehrabi et al. 2021). Second, the physician decides whether to use the tool in a given case and, conditional on use, whether to follow its recommendation. This setup allows us to study whether liability achieves its intended goal by inducing disparity-reducing investment, or instead weakens adoption and shifts investment in ways that can leave disadvantaged patients worse off.

We formalize this interaction as follows. When the algorithm is disparate, following an incorrect recommendation for a disadvantaged patient creates a specific liability exposure; when the algorithm is equal-accuracy, this disparity-contingent liability channel is removed. The physician's core trade-off is then clear: AI can reduce clinical uncertainty, but its use also brings utilization costs and expected legal risk. This structure captures a central regulatory tension: policies enforced at the point of use can feed back to the design stage, in which disparity itself is produced.

Our analysis yields three main results. First, liability tied to adverse outcomes from a disparate AI tool can induce disparate *use* of AI. Even holding the algorithm's accuracy fixed, the physician consults AI for a narrower set of disadvantaged patients because expected legal exposure acts like an additional shadow cost of reliance for that group. A policy designed to shield disadvantaged patients from unequal algorithmic performance can thus end up limiting their access to AI altogether.

Second, the effect of liability on AI use for disadvantaged patients is non-monotone. For small increases in liability, the direct deterrence effect dominates and use declines. As liability rises further, however, the firm responds endogenously by reallocating investment toward disadvantaged-group accuracy and, beyond a threshold, by switching to an equal-accuracy design. These upstream responses reduce expected physician exposure and can restore utilization, or even increase it relative to intermediate-liability levels. This feedback loop means the same policy that suppresses downstream use at low levels can strengthen upstream incentives to reduce disparity at higher levels.

Third, we show mandating equal measured accuracy across groups does not necessarily improve welfare and can, under plausible conditions, reduce welfare for *both* groups. A one-size-fits-all accuracy requirement distorts the firm's resource allocation under asymmetric improvement costs: in many cases, it lowers advantaged-group accuracy substantially while raising disadvantaged-group accuracy only modestly. When combined with reimbursement structures that reward AI use, this shift can also alter clinical deployment in ways that increase inappropriate use. Equalizing algorithmic performance and ensuring appropriate clinical reliance are distinct policy objectives. Liability standards aimed at the former may require complementary instruments—such as reimbursement design, utilization guidelines, or monitoring rules—to achieve the latter.

Growing evidence shows algorithmic performance can vary systematically across groups in ways that materially affect care and outcomes (Abràmoff et al. 2022a, Goodman et al. 2023). For example, Obermeyer et al. (2019) show racial disparity in a widely used clinical risk algorithm led to fewer Black patients being identified for additional care despite greater illness severity. Although addressing disparity early in development is often more effective than retrofitting ex post (Gichoya et al. 2023), our analysis shows regulatory design remains crucial: the same legal objective can generate very different equilibrium outcomes depending on how incentives are structured. We do not argue against regulating algorithmic disparity; rather, we clarify the incentive trade-offs that must be addressed if liability standards are to work as intended.

More broadly, this paper treats medical AI as an input to expert decision-making, not a substitute for it. Even when AI is highly accurate, its deployment hinges on transparency, accountability, and the incentives of the human expert in the loop. Regulations such as § 92.210 constrain reliance on disparate tools, implying that the welfare consequences of AI design are filtered through downstream adoption decisions. Accordingly, our model allows physicians to decide whether to use AI and, when they do, how much to rely on it; we study how these deployment decisions feed back into firms' design incentives, and how liability and reimbursement shape that feedback.

2. Literature

Our work contributes to the expert-service literature originating from Darby and Karni (1973), who argue credence-goods providers such as physicians may overprovide expert services. A key

theme in this literature is that liability shapes clinical behavior: empirical and experimental evidence confirms malpractice liability influences physician decision-making (Currie and MacLeod 2008, Dulleck et al. 2011), and theoretical work shows liability can discipline providers much as reputational concerns do (Fong and Liu 2018) or induce appropriate treatment choices (Chen et al. 2022).[1] The introduction of AI complicates this picture. Dai and Singh (2025), for instance, explore physician behavior under emerging liability frameworks but focus on malpractice liability and on liability from disregarding an AI recommendation, without addressing the potential disparity of AI algorithms across patient groups. We model an AI-specific healthcare regulation and analyze its impact on AI firms, physicians, and patients.

Our work also relates to the literature on the role of AI in medical decision-making. Using AI to support clinical decisions connects to the literature on information acquisition. Prior work has examined how people interact with algorithms, including algorithm preference and algorithm aversion (e.g., Dietvorst et al. 2018, Iyer and Ke 2024, Leung et al. 2018, Mohammadi et al. 2024). The literature has also attempted to identify the causes of overuse/over-adherence of AI algorithms. McLaughlin and Spiess (2022) provide an explanation that AI altering preferences can lead to over-adherence; for instance, a decision maker may view the algorithmic recommendation as a default action. In terms of the cause of AI underuse, Dai and Singh (2020) develop a signaling model to show highly skilled physicians may underutilize diagnostic tests to signal their skills. Balakrishnan et al. (2026) show humans over-adhere to an algorithm's predictions when their own private information, which the algorithm cannot access, is valuable, and under-adhere otherwise. We differ from this stream by not directly investigating the influence of human factors but examining the influence of potential liability. By connecting both the upstream and downstream of medical AI, we show a physician may underuse AI under relatively small liability, due to a dominant liability concern, and overuse AI under relatively large liability, due to its incentive for the upstream to invest in high algorithmic accuracy.

Our work also contributes to the literature on disparate algorithm performance. Extensive research documents systematic differences in health and healthcare by race, gender, age, and other characteristics (e.g., Heckler 1985, Nelson 2002). As predictive algorithms diffuse across high-stakes domains, a central concern is that these tools can reproduce or even widen

[1]The related literature on liability in other domains, such as product safety (Guan et al. 2024, Iyer and Singh 2018), offers valuable insights but does not directly apply to the expert-service context.

such gaps when training data and prediction targets are misaligned with clinical needs (e.g., Benjamin 2016, Gianfrancesco et al. 2018). For example, Tipton et al. (2023) review evidence that standard "fairness fixes" (such as omitting race) can improve parity along one dimension while worsening broader health outcomes. Related work develops technical approaches to reducing disparate performance—by changing objectives (Samorani et al. 2022), redefining targets (Obermeyer et al. 2019), or incorporating group identity (Gillis et al. 2021)—and, in parallel, economic models of algorithm design under data investment, disclosure, and fairness concerns (Diao et al. 2023, Li and Li 2023). We add a complementary mechanism: deployment-facing regulation can narrow measured performance gaps yet induce unequal use in equilibrium; put differently, equalizing algorithmic accuracy need not equalize who benefits from AI.

This paper also contributes to the economics and marketing literature on product design under fairness/ethical or policy constraints (e.g., Diao et al. 2023, Israeli 2018, Iyer and Ke 2024, Ke and Sudhir 2023). As in studies of how marketing instruments (for example, incentives, pricing, trust signals) influence intermediary adoption of new technologies (e.g., Lambrecht and Tucker 2024, Luo et al. 2019, Zimmermann et al. 2024), we study how fairness constraints reshape physician–AI interaction and adoption. Relative to that work, which sets product quality for a buyer who decides whether to purchase, in our paper an expert intermediary, not the buyer, governs use, and we endogenize the quality of the expert's information technology that the credence-service literature takes as given. Equalizing measured quality then need not equalize realized access. Relatedly, our paper engages with the literature on the implications of algorithmic fairness. Corbett-Davies et al. (2017) reveal a tension between improving public safety and satisfying prevailing notions of algorithmic fairness when deciding whether to release pretrial defendants back into the community. Shimao et al. (2022) show fair algorithms can lead to different equilibrium behaviors among different groups of prediction subjects. Corbett-Davies and Goel (2018) find equal opportunity and demographic parity may harm the protected group due to heterogeneity across groups. Liu et al. (2018) show an overly aggressive fairness criterion may cause harm to the protected group in the long term, because giving too many loans to people in a protected group who cannot pay them back can hurt the group's credit scores on average. Fu et al. (2022) show fair algorithms that require impact parity can make everyone worse off, including the protected group, because of the firm's strategic behavior of underinvesting in learning. Our mechanism is distinct: rather than relying on heterogeneity in patient characteristics across groups, we

show mandating equal accuracy can harm disadvantaged patients even when the two groups differ only in the cost of improving algorithmic performance. The welfare loss arises from the physician's dual objective of representing patients while also pursuing AI reimbursement, which distorts usage decisions under a one-size-fits-all accuracy constraint.

3. Model

Consider a physician who selects a treatment for a patient from two possible options, T_1 and T_2. One option is *appropriate* for the patient and the other is *inappropriate*. The benefit of the appropriate treatment for any patient is b, where $b > 0$, whereas the benefit of the inappropriate treatment is 0.

To reflect the health disparities often observed across demographic groups (Obermeyer et al. 2019), we consider two patient types, $t \in \{x, y\}$. Type-x patients represent an advantaged group, such as White patients, whereas type-y patients represent a disadvantaged group, such as Black patients. Following the protected characteristics outlined in Section 1557 of the Affordable Care Act (CMS 2024), we assume a patient's type, t, is observable. For simplicity, we normalize the mass of type-x patients to one and that of type-y patients to β, where $0 < \beta \leq 1$, to capture that disadvantaged patients are often a minority group.

The physician holds a prior belief that treatment T_1 is appropriate for a patient with probability α, whereas treatment T_2 is appropriate with probability $1-\alpha$, where $\alpha \sim U[0,1]$.[2] We interpret α as the physician's case-specific assessment after her usual initial examination of the patient and before she consults the AI; it is not a population base rate, and its distribution captures across-case heterogeneity in baseline diagnostic uncertainty rather than any group-specific bias. We study the case of type-dependent priors in Section 6.3. We use a_t to denote the appropriate treatment for a type-t patient. The physician has access to an AI-powered clinical algorithm (hereafter "AI"). The AI generates a treatment signal, which the physician uses to update her belief about the appropriate treatment. We denote the AI signal for a type-t patient by s_t and define AI accuracy for type t as ρ_t, where $t = x, y$. Specifically, $P_{1|1} := \mathcal{P}(s_t = T_1 | a_t = T_1) = \rho_t$ and $P_{2|1} := \mathcal{P}(s_t = T_2 | a_t = T_1) = 1 - \rho_t$. Similarly, $P_{2|2} := \mathcal{P}(s_t = T_2 | a_t = T_2) = \rho_t$ and $P_{1|2} := \mathcal{P}(s_t = T_1 | a_t = T_2) = 1 - \rho_t$.

[2]The assumption of type-independent priors helps isolate the incremental role of algorithmic disparity and the specific anti-discrimination liability channel we study in this paper. Holding the physician's priors the same for both patient types ensures any cross-group differences in outcomes in our analysis arise from the AI's differential accuracy and the physician's endogenous AI-use decision under liability.

AI recommendations are informative but imperfect, that is, $1/2 < \rho_t < 1$. The physician observes the algorithm's accuracy for each patient type, ρ_t, through her clinical expertise and experience with the tool. We call an algorithm *equal-accuracy* if $\rho_x = \rho_y$, and in that case we write the common accuracy as ρ (dropping the subscript t). An algorithm with $\rho_x > \rho_y$ delivers lower accuracy for type-y patients and is therefore disparate for type-y patients.

With the AI signal, the physician updates her belief about the appropriate treatment using Bayes' rule. The physician's posterior belief that T_1 is appropriate is

$$Q_{1|1} := \mathcal{P}(a_t = T_1 | s_t = T_1) = \frac{\alpha\rho_t}{\alpha\rho_t + (1-\alpha)(1-\rho_t)},$$
$$Q_{1|2} := \mathcal{P}(a_t = T_1 | s_t = T_2) = \frac{\alpha(1-\rho_t)}{\alpha(1-\rho_t) + (1-\alpha)\rho_t}.$$

The physician believes T_2 to be appropriate with the complementary probability, that is, $Q_{2|1} := \mathcal{P}(a_t = T_2 | s_t = T_1) = 1 - Q_{1|1}$ and $Q_{2|2} := \mathcal{P}(a_t = T_2 | s_t = T_2) = 1 - Q_{1|2}$.

Consistent with the Section 1557 clinical algorithm provision (§ 92.210), we model deployment-facing liability as follows. If the physician follows the signal of a disparate clinical algorithm for a type-y patient and the resulting treatment is inappropriate, she faces a liability cost ℓ (CMS 2022). This channel is specific to type-y patients in our setting; we abstract from other sources of legal exposure (for example, malpractice) to isolate the incentive effects of the provision. Reflecting that the provision governs the use of decision-support tools and applies to covered entities rather than to developers (Mello and Roberts 2024), we assign liability to the physician rather than to the AI firm. When the algorithm has equal accuracy across patient types ($\rho_x = \rho_y$), this disparity-triggered liability channel does not apply.

When the physician deploys the AI tool in a patient's case, the patient incurs a cost $c > 0$. We model c as a reduced-form burden that can include out-of-pocket expenses and non-monetary disutility—for example, psychological discomfort, perceived risk, or distrust when an algorithm is involved in diagnosis or treatment (Longoni et al. 2019). Importantly, c is not meant to give patients control over AI use: although opt-out is feasible in some settings, clinical practice and accountability assign the deployment decision to the physician, and we therefore assume the physician retains full discretion over whether to use the tool.

We also allow the physician to receive a per-use benefit $r > 0$ when she deploys the tool. This term captures reimbursement for AI-assisted services—for example, payment pathways enabled by AI-specific Current Procedural Terminology (CPT) codes and New

Technology Add-On Payments (NTAP)—and broader private returns from using AI (Parikh and Helmchen 2022). Depending on the setting, r may also reflect reputational or professional gains from being perceived as technologically capable or clinically sophisticated (Liaw et al. 2022, Schubert et al. 2025, Schuitmaker et al. 2025).[3,4]

When the physician decides whether and how to use AI, she cares about both the patient's health outcome (that is, treatment benefit net of AI-use cost) and her own nonclinical objectives (that is, revenue and liability). We assume the physician assigns a weight $\theta \geq 0$ to her nonclinical objectives. A physician with $\theta = 0$ is altruistic and cares only about the patient's health outcome, whereas a physician with $\theta > 0$ is *impurely* altruistic and also cares about her nonclinical objectives. Unless otherwise noted, we assume $\theta > 0$; the altruistic case $\theta = 0$ serves as a benchmark for comparison.

Next, we introduce a profit-maximizing firm that supplies an AI-powered clinical algorithm with accuracy ρ_t for patient type $t \in \{x, y\}$. Consistent with fee-for-service arrangements in which AI use is reimbursed at predetermined rates, we assume the firm receives a payment f each time the physician deploys the tool in a patient's case (Abràmoff et al. 2022b, 2024).[5] Achieving accuracy ρ_t for type t entails a development cost $\kappa_t(\rho_t - \frac{1}{2})^2$, where $\kappa_t > 0$ is type-specific. To capture the limited availability and higher acquisition cost of training data for disadvantaged patients, we assume $\kappa_x < \kappa_y$. We focus on settings in which the firm trains a new model on a fixed historical dataset, rather than incrementally updating an existing system, and thus treat the cost coefficients as exogenous and independent of downstream usage. The cost asymmetry $\kappa_x < \kappa_y$ reflects the empirically common case in which improving performance for the disadvantaged group is harder.[6] The separable quadratic

[3]Several U.S. providers now charge patients directly for AI-assisted reads; for example, RadNet's imaging centers offer an optional AI review of a screening mammogram for an out-of-pocket fee, which a substantial share of patients elect, illustrating a patient-side cost $c > 0$ and, where the provider bills for the read, a per-use revenue $r > 0$. On the patient side, $c > 0$ is also consistent with documented resistance to medical AI (Longoni et al. 2019).

[4] In the special case in which $r < 0$, possibly due to a strong negative reputational effect of using AI, our main results on the physician's reduced AI use for disadvantaged patients and the non-monotone effect of liability on the physician's AI use for disadvantaged patients continue to hold. However, in this case, disadvantaged patients are better off, whereas advantaged patients are worse off as a result of the equal-accuracy mandate. Because the physician has no incentive to overuse AI in the $r < 0$ case, the disadvantaged patients cannot be worse off.

[5]In a model extension presented in Section 6.1, we assume the AI firm sets a profit-maximizing price f and find all our main insights continue to hold qualitatively under the endogenous pricing assumption.

[6]If the asymmetry were reversed ($\kappa_y < \kappa_x$), the algorithmic accuracy for the disadvantaged group would not be harder to improve and the disparity-contingent liability channel would be weak or inactive; in an application in which the type-y group is the lower-cost group, the labels x and y can be exchanged.

structure is a parsimonious way to capture the practical fact that improving performance for a harder-to-predict subgroup typically requires targeted effort (for example, subgroup-specific tuning or additional data), even when the deployed model is unified.

Given physician demand $(d_x^{\mathrm{D}}, d_y^{\mathrm{D}})$ under a disparate design, the firm's profit is

$$\pi^{\mathrm{D}}(\rho_x, \rho_y) := (d_x^{\mathrm{D}} + d_y^{\mathrm{D}})f - \kappa_x(\rho_x - \tfrac{1}{2})^2 - \kappa_y(\rho_y - \tfrac{1}{2})^2, \tag{1}$$

where d_x^{D} and d_y^{D} denote the volumes of type-x and type-y cases in which the physician deploys the tool (expressions are given in eq. (6)). When the firm supplies an equal-accuracy design, $\rho_x = \rho_y \equiv \rho$, profit can be written as

$$\pi^{\mathrm{E}}(\rho) := (d_x^{\mathrm{E}} + d_y^{\mathrm{E}})f - \kappa_x(\rho - \tfrac{1}{2})^2 - \kappa_y(\rho - \tfrac{1}{2})^2, \tag{2}$$

where $(d_x^{\mathrm{E}}, d_y^{\mathrm{E}})$ are the corresponding demands under equal accuracy, given in eq. (7).

We assume $c > \theta r$, which eliminates cases in which the AI firm develops an arbitrarily bad algorithm, but the physician still uses AI for all patients. In other words, we do not consider cases in which the physician uses AI solely for private gain from reimbursement. In addition, although parameters c, r, and f are likely interrelated in practice, they do not reflect a direct transfer of funds. Instead, funds typically flow from patients and insurers to providers (physician, hospital, or health system), who then pay AI firms. For clarity, we summarize the meaning, examples, and payment flow for parameters c, r, and f in Table 1. To maintain tractability and focus on core strategic trade-offs, our baseline model treats parameters c, r, and f as independent. Nonetheless, in Section 6.1, we explore an alternative model in which the patient's AI-use cost is the same as the firm's price ($c = f$), and we confirm our main results remain robust under this specification.

Next, we describe the patient's expected utility. First, consider the case in which the physician decides not to use AI. In this case, the physician follows her own prior belief to treat the patient. Given the physician's prior belief α of treatment T_1 being appropriate, the patient's expected utility from treatment T_1 is αb, and from treatment T_2 is $(1-\alpha)b$. Now consider the case in which the physician uses AI. In this case, the patient incurs an AI-use cost c. The patient's expected benefit from treatment depends on the AI signal and the physician's treatment choice. Specifically, if the physician prescribes treatment T_j after observing signal T_i, the patient's expected benefit is $Q_{j|i}b$, where $i, j \in \{1, 2\}$. We assume the

Table 1: Meaning, Examples, and Payment Flow for Parameters c, r, and f

	Meaning	**Example**	**Payment Flow**
c	Patient's cost of AI use	Out-of-pocket expenses (copayments, deductibles, coinsurance) or disutility (distrust, privacy concerns)	Borne by patient
r	Provider's per-use revenue	Insurer reimbursement (for example, CMS, private insurers) and reputational benefit, if any, from AI use	Accrued by provider
f	AI firm's per-use fee	Payment from provider, hospital, or health system to AI firm for technology use or licensing	Provider $\rightarrow$ AI firm

Notes. The parameter c captures both monetary costs (for example, copayments) and non-monetary disutility borne by the patient. The parameter r captures both financial reimbursement (for example, CMS payments via CPT codes) and non-monetary gains such as perceived competence or technological savviness. Only f represents a direct inter-party transfer.

patient's expected utility is determined by her realized health outcome and any incurred cost. We do not include any gain from lawsuit payouts, to avoid unrealistic scenarios in which the physician might intentionally harm the patient to generate a financial transfer to the patient. Equivalently, expected patient welfare for each type, denoted W_x and W_y and defined formally in Section 5.3, measures patient health surplus (expected treatment benefit net of the AI-use cost) and excludes litigation compensation. If litigation compensation enters only as an ex post accounting transfer that no decision maker internalizes, equilibrium behavior is unaffected, because patients choose neither AI use nor litigation effort and the provider's exposure is already summarized by ℓ; the transfer then raises compensation-inclusive type-y surplus and cancels in total surplus apart from deadweight legal costs. If instead the physician internalizes the patient's compensation through patient utility, the behavioral model changes, and we analyze that formulation in Section 5.3, where the weight on the liability term becomes $\theta - 1$. Our welfare comparisons should therefore be read as comparisons of ex ante health surplus. The patient's expected utility is summarized in Table 2.

We now describe the physician's expected payoff from treating an individual patient. If the physician uses AI, she receives a per-use benefit r and may face liability. For type-y patients, liability arises only when the physician follows the AI recommendation and that recommendation is incorrect. Accordingly, following signal T_1 leads to liability with probability $Q_{2|1}$, whereas following signal T_2 leads to liability with probability $Q_{1|2}$. We

Table 2: The Patient's Expected Utility When the Physician Uses AI

	Physician's decision	
AI signal	T_1	T_2
$s_t = T_1$	$Q_{1\|1} \cdot b - c$	$Q_{2\|1} \cdot b - c$
$s_t = T_2$	$Q_{1\|2} \cdot b - c$	$Q_{2\|2} \cdot b - c$

interpret ℓ as a reduced-form expected exposure borne by the provider rather than a literal per-case statutory fine. Section 92.210 does not impose strict liability for every adverse outcome; it prohibits discriminatory use of decision-support tools and requires covered entities to make reasonable efforts to identify and mitigate disparity. Accordingly, ℓ captures the expected cost of scrutiny, enforcement, settlement, and remediation conditional on relying on a disparate tool in a type-y case that yields an inappropriate treatment (Mello and Roberts 2024). Because the physician finds following the signal optimal whenever she uses AI, this event is equivalently described in terms of the realized treatment. The physician values both the patient's health outcome net of cost and her own nonclinical payoff, placing weight θ on the latter. Table 3 summarizes the physician's payoff, where $\mathbf{1}_{t=y}$ equals 1 if $t = y$ and 0 otherwise.

Table 3: The Physician's Expected Payoff When Treating an Individual Patient Using AI

	Physician's decision	
AI signal	T_1	T_2
$s_t = T_1$	$Q_{1\|1} \cdot b - c + \theta(r - Q_{2\|1} \cdot \mathbf{1}_{t=y}\,\ell)$	$Q_{2\|1} \cdot b - c + \theta r$
$s_t = T_2$	$Q_{1\|2} \cdot b - c + \theta r$	$Q_{2\|2} \cdot b - c + \theta(r - Q_{1\|2} \cdot \mathbf{1}_{t=y}\,\ell)$

Figure 1 summarizes the timing. First, the AI firm chooses whether to offer a disparate design or an equal-accuracy design and selects the corresponding accuracy level(s) ρ_t for each patient type. Second, the physician decides whether to use AI for a given patient. If she uses AI, the algorithm generates a treatment signal and the physician then decides whether to follow it; if she does not use AI, she selects treatment based only on her prior belief. Finally, the patient's health outcome is realized and provider liability is assessed under the clinical

algorithm provision. We solve the game by backward induction. Throughout, we assume $\kappa_y < \frac{f\beta(b+\theta\ell)^2}{b(2c-2\theta r+\theta\ell)}$, which ensures that, in equilibrium, the physician uses AI for a strictly positive measure of type-y patients (that is, $d_y^{\mathrm{D}*} > 0$).

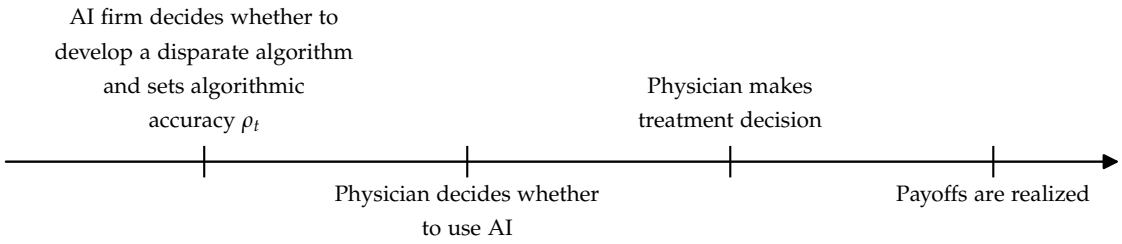


Figure 1: Timing of the Game. The AI firm's first-stage choice is between a disparate design ($\rho_x > \rho_y$) and an equal-accuracy design ($\rho_x = \rho_y$).

4. Analysis

In this section, we first analyze in Section 4.1 the physician's decision of whether and how to use AI, taking AI accuracy as given. We then analyze in Section 4.2 how the AI firm chooses AI accuracy.

4.1 Downstream Implications

We analyze the physician's decision of whether to use a disparate AI algorithm with $\rho_x > \rho_y > 1/2$ for type-x and type-y patients. The ordering $\rho_x > \rho_y$ here defines the disparate-design *subgame* whose downstream behavior we characterize; it is not an equilibrium restriction. Proposition 1 shows that whenever the firm finds a disparate design optimal, $\rho_x^* > \rho_y^*$ emerges endogenously, so the ordering is derived rather than assumed. Note that the analysis of the physician's decision to use an equal-accuracy algorithm ($\rho_x = \rho_y > 1/2$) is analogous to the analysis for type-x patients under a disparate algorithm.

If the physician does not use AI, then by comparing the physician's payoff from prescribing treatment T_1 (that is, αb) and T_2 (that is, $(1-\alpha)b$), we know the physician prescribes treatment T_1 if $\alpha > 1/2$ and prescribes treatment T_2 if $\alpha \leq 1/2$. We can summarize

the physician's expected payoff when she does not use AI as follows:

$$U = \begin{cases} (1-\alpha)b, & \text{if } \alpha \leq 1/2 \\ \alpha b, & \text{if } \alpha > 1/2. \end{cases} \tag{3}$$

Now suppose the physician uses AI. In this case, the probability of the AI signal suggesting treatment T_1 is $\mathcal{P}(s_t = T_1) = \alpha \cdot \rho_t + (1-\alpha) \cdot (1-\rho_t)$ and treatment T_2 is $\mathcal{P}(s_t = T_2) = \alpha \cdot (1-\rho_t) + (1-\alpha) \cdot \rho_t$. Next, we analyze the physician's decision of whether to use AI.

4.1.1 Type-x Patient

Suppose the physician consults AI for a type-x patient. If the signal is $s_x = T_1$, she compares the payoffs from prescribing T_1 and T_2, namely $Q_{1|1}b - c + \theta r$ and $Q_{2|1}b - c + \theta r$ (see Table 3). She therefore prescribes T_1 if and only if $\alpha > 1 - \rho_x$, and prescribes T_2 otherwise. If the signal is $s_x = T_2$, she compares $Q_{1|2}b - c + \theta r$ and $Q_{2|2}b - c + \theta r$, and prescribes T_1 if and only if $\alpha > \rho_x$, and T_2 otherwise. Because $1/2 < \rho_x < 1$, we have $1 - \rho_x < \rho_x$, so the prior space is partitioned into three regions. The physician's expected payoff from AI use for a type-x patient is

$$U_x = \begin{cases} \mathcal{P}(s_x = T_1)Q_{2|1}b + \mathcal{P}(s_x = T_2)Q_{2|2}b - c + \theta r, & \text{if } \alpha \leq 1 - \rho_x, \\ \mathcal{P}(s_x = T_1)Q_{1|1}b + \mathcal{P}(s_x = T_2)Q_{2|2}b - c + \theta r, & \text{if } 1 - \rho_x < \alpha \leq \rho_x, \\ \mathcal{P}(s_x = T_1)Q_{1|1}b + \mathcal{P}(s_x = T_2)Q_{1|2}b - c + \theta r, & \text{if } \alpha > \rho_x. \end{cases} \tag{4}$$

Next, we present the physician's decision of whether and how to use AI for type-x patients. A comparison of the physician's expected payoff when using AI and not using AI for type-x patients reveals the following lemma. (All proofs are in the Appendix.)

Lemma 1. *For type-x patients, the physician uses AI if and only if*

$$\frac{(1-\rho_x)b + c - \theta r}{b} < \alpha < \frac{\rho_x b - c + \theta r}{b}.$$

Conditional on using AI, the physician follows the AI signal.

The physician's prior belief α represents the probability that treatment T_1 is appropriate before observing any AI recommendation. Without AI, the physician follows this prior and

prescribes T_1 when $\alpha > 1/2$. When AI is used, the physician updates her belief using Bayes' rule, and the signal realization moves her posterior toward or away from T_1 by an amount that increases in the accuracy ρ_x. The use of AI creates value only when its signal can change the physician's treatment decision.

Intuitively, when α is very low, the physician already believes T_2 is likely appropriate, and AI's informational value cannot justify the additional patient cost c. When α is very high, the physician is confident in T_1 and does not expect AI to improve her decision. Only when α lies in the intermediate range does AI meaningfully reduce uncertainty, making its use optimal. Moreover, because AI is informative ($1/2 < \rho_x < 1$), whenever the physician chooses to use it, she rationally follows its recommendation.

4.1.2 Type-y Patient

Now suppose the physician uses AI to generate the signal for a type-y patient. If the AI signal is T_1, the physician compares her payoffs from prescribing treatments T_1 and T_2, which are $Q_{1|1} \cdot b - Q_{2|1} \cdot \theta\ell - c + \theta r$ and $Q_{2|1} \cdot b - c + \theta r$ (see Table 3), respectively, and prescribes treatment T_1 if $\alpha > \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell}$ and treatment T_2 if $\alpha \leq \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell}$. However, if the AI signal is T_2, by comparing her payoffs from prescribing treatments T_1 and T_2, which are $Q_{1|2} \cdot b - c + \theta r$ and $Q_{2|2} \cdot b - Q_{1|2} \cdot \theta\ell - c + \theta r$ (see Table 3), respectively, the physician prescribes treatment T_1 if $\alpha > \frac{b\rho_y}{b+(1-\rho_y)\theta\ell}$ and treatment T_2 if $\alpha \leq \frac{b\rho_y}{b+(1-\rho_y)\theta\ell}$. Given the assumption that the physician uses AI for at least some type-y patients, we have $\rho_y > \frac{b+2c+2\theta\ell-2\theta r}{2b+2\theta\ell}$. In addition, because $\frac{b+2c+2\theta\ell-2\theta r}{2b+2\theta\ell} > \frac{b+\theta\ell}{2b+\theta\ell}$, it follows that $\rho_y > \frac{b+\theta\ell}{2b+\theta\ell}$. It is straightforward that $\frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} < \frac{1}{2} < \frac{b\rho_y}{b+(1-\rho_y)\theta\ell}$. Therefore, if the physician uses AI for a type-y patient, her expected payoff is given by

$$U_y = \begin{cases} \mathcal{P}(s_y = T_1) \cdot Q_{2|1} \cdot b + \mathcal{P}(s_y = T_2) \cdot \left(Q_{2|2} \cdot b - Q_{1|2}\theta\ell\right) - c + \theta r, & \text{if } \alpha \leq \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} \\ \mathcal{P}(s_y = T_1) \cdot \left(Q_{1|1} \cdot b - Q_{2|1}\theta\ell\right) + \mathcal{P}(s_y = T_2) \cdot \left(Q_{2|2} \cdot b - Q_{1|2}\theta\ell\right) - c + \theta r, & \text{if } \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} < \alpha \leq \frac{b\rho_y}{b+(1-\rho_y)\theta\ell} \\ \mathcal{P}(s_y = T_1) \cdot \left(Q_{1|1} \cdot b - Q_{2|1}\theta\ell\right) + \mathcal{P}(s_y = T_2) \cdot Q_{1|2} \cdot b - c + \theta r, & \text{otherwise.} \end{cases} \quad (5)$$

The following lemma describes the physician's decision of whether to use AI for type-y patients.

Lemma 2. *For type-y patients, the physician uses AI if and only if*

$$\frac{(1-\rho_y)(b+\theta\ell) + c - \theta r}{b} < \alpha < \frac{\rho_y b - (1-\rho_y)\theta\ell - c + \theta r}{b}.$$

Conditional on using AI, the physician follows the AI signal.

As in the type-x case, the physician uses AI for type-y patients only when the informational value of consultation is sufficiently high. Equivalently, AI use occurs over an intermediate range of α, in which clinical uncertainty about the appropriate treatment is greatest. For an existing algorithm (that is, exogenous ρ_y), the liability rule unambiguously makes AI less attractive for type-y patients. Holding ρ_y fixed, an increase in liability ℓ contracts the set of priors for which AI is used.

Because liability is triggered when a physician follows an erroneous recommendation from a disparate algorithm in type-y cases, one might expect strategic rejection of AI advice. Our model shows otherwise: conditional on consulting AI, the physician optimally follows the signal. The intuition is straightforward. If she consults AI but ignores the signal regardless of its realization, she incurs the consultation cost c without obtaining informational benefit; with $c > \theta r$, she is strictly better off not consulting at all. If she follows only when the signal is T_1 (or only when it is T_2), her behavior is equivalent to always choosing T_1 (or always choosing T_2), which is again weakly dominated by making that fixed treatment choice without AI. Hence, when the physician uses AI for type-y patients, she follows the AI signal.

Next, we examine how the regulation affects the physician's relative use of AI for type-x and type-y patients. Comparing the AI-use intervals in Lemmas 1 and 2 shows that, in the presence of liability, the physician is (weakly) less likely to use AI for type-y patients than for type-x patients.

Under the Section 1557 clinical algorithm provision, a physician is exposed to liability when her reliance on a disparate AI tool leads to an inappropriate treatment for a type-y patient. This liability functions as an added expected cost of AI use for type-y cases, making adoption and reliance more selective for that group—even when the same tool is available for type-x patients. By contrast, type-x cases do not activate this liability channel; therefore, AI use for type-x patients is governed only by the accuracy–cost–incentive trade-off in Lemma 1. The resulting asymmetry in expected liability generates unequal utilization in equilibrium, with lower AI use among disadvantaged patients—the very group the policy is intended to protect.

4.2 Upstream AI Firm's Accuracy Decision

Anticipating the physician's downstream use of the algorithm, the AI firm chooses whether to offer a disparate design or an equal-accuracy design, and selects the associated accuracy levels ρ_t^* for each patient type. By Lemmas 1 and 2, under a disparate algorithm, the volumes of type-x and type-y patients for whom the physician uses AI are

$$d_x^{\mathrm{D}} = \frac{(2\rho_x - 1)b - 2c + 2\theta r}{b} \text{ and } d_y^{\mathrm{D}} = \beta \cdot \frac{(2\rho_y - 1)b - 2(1-\rho_y)\theta\ell - 2c + 2\theta r}{b}. \tag{6}$$

Here $0 < \beta \leq 1$ scales the mass of type-y patients relative to type-x patients, and the superscript D denotes the disparate design. The firm then chooses (ρ_x, ρ_y) to maximize expected profit in eq. (1).

Next, consider the equal-accuracy design, under which the algorithm attains the same accuracy for both patient types, $\rho_x = \rho_y \equiv \rho$. The resulting volumes of AI use for type-x and type-y patients are

$$d_x^{\mathrm{E}} = \frac{(2\rho - 1)b - 2c + 2\theta r}{b} \text{ and } d_y^{\mathrm{E}} = \beta \cdot \frac{(2\rho - 1)b - 2c + 2\theta r}{b}, \tag{7}$$

where the superscript E denotes the equal-accuracy design. The firm then chooses ρ to maximize expected profit in eq. (2).

For ease of presentation, we define a threshold $\widetilde{\ell}$ such that the AI firm's expected profit for a disparate algorithm (which is strictly decreasing in ℓ) exceeds its expected profit for an equal-accuracy algorithm when $\ell < \widetilde{\ell}$. Otherwise, the AI firm develops an equal-accuracy algorithm in equilibrium. The following proposition presents the equilibrium AI accuracy set by the firm.

Proposition 1. *Suppose* $\beta(\beta+2)\kappa_x > \kappa_y$ *and* $\max\{\frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)}, \frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2}\} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y}{2\beta(b+\theta\ell)}\}$. *There exists a unique cutoff* $\widetilde{\ell}$ *in the feasible region such that:*

(a) if $\ell < \widetilde{\ell}$, *the AI firm develops a disparate algorithm with* $\rho_x^* = \frac{1}{2} + \frac{f}{\kappa_x}$ *and* $\rho_y^* = \frac{1}{2} + \frac{\beta f(b+\theta\ell)}{b\kappa_y}$, *and* $\rho_x^* > \rho_y^*$;

(b) if $\ell \geq \widetilde{\ell}$, *the AI firm develops an equal-accuracy algorithm with* $\rho^* = \frac{1}{2} + \frac{(1+\beta)f}{\kappa_x+\kappa_y}$.

Recall that the AI firm earns a per-use fee f whenever the physician consults AI and incurs cost $\kappa_t(\rho_t - \frac{1}{2})^2$ to deliver accuracy ρ_t for type-t patients. Because improving accuracy

for type-y patients is more costly ($\kappa_y > \kappa_x$) and demand from that segment is smaller ($\beta \leq 1$), the firm optimally allocates more performance to type-x patients in the disparate regime, so $\rho_x^* > \rho_y^*$.

As liability ℓ rises, physician demand for AI in type-y cases becomes more sensitive to expected legal exposure, reducing the profitability of a disparate design. Once ℓ exceeds $\widetilde{\ell}$, the firm optimally switches to an equal-accuracy design, which removes the disparity-triggered liability channel and helps preserve overall use. Although ρ_y^* increases with ℓ within the disparate regime, it does not overtake ρ_x^*; before that point, the firm prefers to switch to an equal-accuracy design.

Proposition 1 highlights a key trade-off: stronger liability pushes design toward fairness, but potentially at the cost of aggregate efficiency in accuracy investment. Weak liability sustains a profit-driven disparate design, whereas strong liability induces convergence to equal accuracy.

5. Managerial and Policy Insights

We now examine how the liability rule influences AI-use disparity, appropriate AI use, and patient welfare.

5.1 AI-Use Disparity

Absent differential costs or constraints, one would expect physicians to rely on AI at similar rates across patient types. In practice, however, AI use can differ systematically across groups. We refer to such differences as *AI-use disparity*—the gap between the share of type-x cases and the share of type-y cases in which the physician uses AI.

AI-use disparity may arise under the Section 1557 rule. Importantly, under certain conditions, the rule may fail to mitigate disparities originating from algorithm development and may instead reduce physicians' incentives to use AI for disadvantaged patients. As a result, the patients whom the rule is designed to protect may be less likely to benefit from AI-assisted decision-making. This raises a key question: when does the rule exacerbate AI-use disparity, and under what conditions can such disparity be narrowed? The following proposition provides insight into this question. For ease of presentation, we define a threshold f_l in the proof of Proposition 2.

Proposition 2. *Suppose* $\beta(\beta+2)\kappa_x > \kappa_y$ *and* $f_l < f < \min\left\{\frac{\kappa_x}{2}, \frac{\kappa_y}{4\beta}\right\}$. *Then, type-y AI use is non-monotone in liability: as* ℓ *increases, the physician uses AI for fewer type-y patients when* $\ell < \frac{b(\kappa_y - 4\beta f)}{4\theta\beta f}$ *and for more type-y patients when* $\frac{b(\kappa_y - 4\beta f)}{4\theta\beta f} \leq \ell < \widetilde{\ell}$. *When* $\ell > \widetilde{\ell}$, *the firm switches to an equal-accuracy design, in which case type-y AI use does not depend on* ℓ.

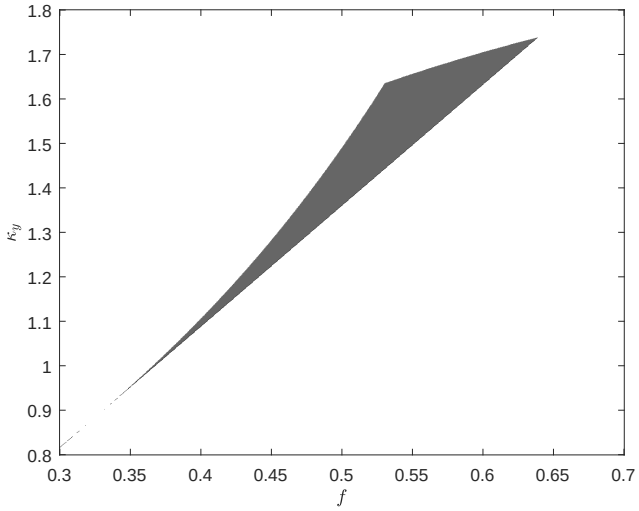


Figure 2: Parameter region in which type-y AI use is non-monotone in liability (Proposition 2).

Since CMS solicited comments on the proposed Section 1557 clinical algorithm provision in August 2022, a recurring concern has been that an increase in liability ℓ could induce physicians to pull back from clinical algorithms and thereby widen disparities in care (see, e.g., Goodman et al. 2023). Proposition 2 sharpens this point: an increase in liability ℓ can either exacerbate or mitigate AI-use disparity, depending on its level. The mechanism is a tug-of-war between two forces. Downstream, higher liability discourages the physician's AI use for type-y patients by directly raising the physician's expected liability cost. Upstream, higher liability increases the firm's incentive to improve accuracy for type-y patients, which makes AI more attractive to the physician and can offset (or reverse) the deterrence effect.

When liability ℓ is small, an increase in its level causes a limited reduction in the physician's AI use for type-y patients. In this case, the firm has a small incentive to invest in type-y accuracy. Therefore, the direct deterrence channel dominates, and the physician's AI use for type-y patients falls as ℓ rises, thus amplifying the AI-use disparity. When liability ℓ is large, the contraction in the physician's AI use for type-y patients becomes salient for the firm, prompting greater investment in type-y accuracy. Because higher accuracy reduces the likelihood of inappropriate treatment and therefore reduces the physician's expected liability cost, the physician's AI use for type-y patients can increase as ℓ rises, narrowing AI-use

disparity.

The bounds on f in Proposition 2 delimit the range over which the two forces are comparable in magnitude. When f is small, the per-use revenue does not justify the higher cost of type-y accuracy ($\kappa_x < \kappa_y$), and the firm supplies an equal-accuracy design, under which type-y use does not vary with ℓ. When f is large, ρ_y^* responds strongly to ℓ, and the resulting accuracy gains outweigh the deterrence effect at every liability level, so type-y use increases monotonically in ℓ. Only over the intermediate range does the firm both prefer a disparate design and respond moderately enough for deterrence to dominate at low liability, which is what produces the turning point. Figure 2 shows the parameter region in which non-monotone AI use occurs.

5.2 (In)Appropriate AI Use

Consider an altruistic physician whose only concern is a patient's health outcome and costs (that is, $\theta = 0$). From eqs. (4) and (5), the expected payoff of an altruistic physician who uses AI for an individual type-t patient is given as follows:

$$U^{\text{al}} = \begin{cases} \mathcal{P}(s_t = T_1) \cdot Q_{2|1} \cdot b + \mathcal{P}(s_t = T_2) \cdot Q_{2|2} \cdot b - c, & \text{if } \alpha \leq 1 - \rho_t \\ \mathcal{P}(s_t = T_1) \cdot Q_{1|1} \cdot b + \mathcal{P}(s_t = T_2) \cdot Q_{2|2} \cdot b - c, & \text{if } 1 - \rho_t < \alpha \leq \rho_t \\ \mathcal{P}(s_t = T_1) \cdot Q_{1|1} \cdot b + \mathcal{P}(s_t = T_2) \cdot Q_{1|2} \cdot b - c, & \text{otherwise} \end{cases} \tag{8}$$

A comparison of the altruistic physician's expected payoffs from following and disregarding AI's recommendation leads to the following lemma.

Lemma 3. *Suppose $c < (\rho_t - \frac{1}{2})b$ for $t \in \{x, y\}$. For type-t patients, an altruistic physician uses AI if and only if*

$$\frac{(1 - \rho_t)b + c}{b} < \alpha < \frac{\rho_t b - c}{b}.$$

Conditional on using AI, the physician follows the AI signal.

We define a physician's AI use for a specific patient type t as *appropriate* if it matches the behavior of an altruistic physician (with $\theta = 0$). If the physician uses AI for more patients than an altruistic physician would, we refer to this as *overuse* of AI; if the physician uses AI for fewer patients, we refer to it as *underuse* of AI. The following proposition compares AI use under the two physician types. We define a threshold ℓ_s (expression provided in

the Appendix) at which the physician's AI use for type-y patients switches from overuse to underuse.

Proposition 3. *The physician's use of AI relative to the clinically appropriate benchmark satisfies the following.*

(a) When the AI firm supplies an equal-accuracy algorithm, the physician may overuse AI for both patient types.

(b) Suppose the interior disparate design is optimal at ℓ, with $\pi_y(\rho_y^*) > 0$ and $0 < r/\ell < 1/2$. If $\frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y} = \frac{r}{\ell}$, then $\rho_y^* = 1 - \frac{r}{\ell}$ and the physician uses AI appropriately for type-y patients. In this case, $\rho_x^* = \frac{1}{2} + \frac{f}{\kappa_x}$ and the physician overuses AI for type-x patients.

(c) Suppose $\pi_y(\rho_y^*) > 0$ and a disparate design is optimal at $\ell = \ell_s$ (specifically, conditions (A4) and (A5) in the Appendix are satisfied). If $\beta(\beta+2)\kappa_x > \kappa_y$ and $\frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y+4\kappa_y\theta r}{2b\beta} - \frac{\kappa_y\sqrt{2b\theta r+4\theta^2 r^2}}{b\beta}\}$, as liability ℓ increases from a low level, the physician's AI use for type-y patients is non-monotone: she overuses AI for small ℓ, underuses AI for intermediate ℓ, and overuses AI again for large ℓ.[7]

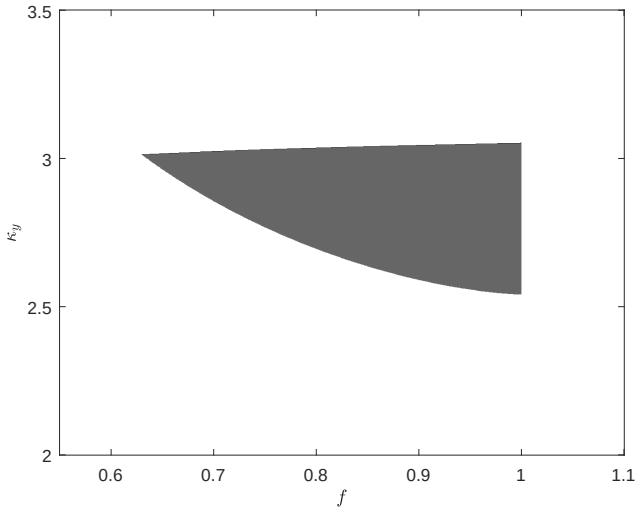


Figure 3: Parameter region in which type-y AI use switches between overuse and underuse (Proposition 3(c)).

Proposition 3(a) highlights a basic tension. An equal-accuracy design removes the type-y-specific liability channel in our setup, but it does not remove the physician's private

[7] Figure 3 provides a numerical illustration of the parameter region in which tri-phase non-monotone AI use occurs.

incentive to use the tool when use is rewarded. When reimbursement r remains in place, the physician's marginal calculus can tilt toward reliance even when AI is not clinically warranted, leading to overuse for both patient types. The broader point is that equalizing algorithmic performance does not, by itself, align deployment incentives; if liability no longer disciplines use under equal accuracy, reimbursement can become the dominant force shaping utilization. This suggests a role for complementary instruments that better tie payment and accountability to clinical value.

Proposition 3(b) shows appropriate use for disadvantaged patients can arise even when the supplied algorithm has unequal accuracy across groups, provided reimbursement exactly offsets the physician's expected liability exposure from relying on AI in type-y cases. The margin condition is $r = (1 - \rho_y^*)\ell$. Using the equilibrium expression for ρ_y^*, this is equivalent to $\frac{r}{\ell} = \frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y}$, which characterizes the combinations of liability and reimbursement that restore appropriate use for type-y patients.

We illustrate Proposition 3(c) in Figure 4. The solid lines plot the belief cutoffs in α that delimit when the physician uses AI for type-y patients; the dash-dot lines report the corresponding cutoffs for an altruistic physician ($\theta = 0$). As liability ℓ rises, equilibrium use for type-y patients can be non-monotone: overuse at low ℓ, underuse at intermediate ℓ, and overuse again when ℓ is large. This pattern contrasts with the view that liability primarily discourages reliance and leads clinicians to abandon AI (see, e.g., Goodman et al. 2023). The mechanism is a shifting balance between downstream deterrence and upstream design responses. For small ℓ, reimbursement dominates expected liability, so the physician relies on the tool too often. For intermediate ℓ, liability becomes first-order and suppresses reliance, generating underuse. For sufficiently large ℓ, the firm is induced to supply an equal-accuracy design, which relaxes the liability channel for type-y patients and restores the reimbursement-driven incentive to use AI, again producing overuse. Notably, in the low-ℓ region in which the physician overuses a disparate tool for type-y patients, increasing ℓ can widen utilization differences across patient types even as the policy is intended to protect the disadvantaged group. The lesson is that equalizing measured performance is not an appropriate-use guarantee; policy must also address the incentives governing clinicians' reliance decisions.

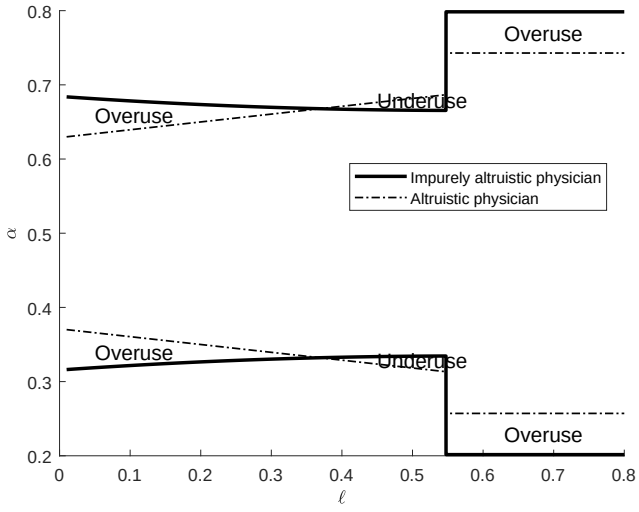


Figure 4: AI use for type-y patients: for a given ℓ, the physician uses AI when α lies between the cutoffs. Dash-dot lines correspond to an altruistic physician ($\theta = 0$); solid lines correspond to an impurely altruistic physician.

5.3 Effect of Mandating Equal Accuracy

Recent policy initiatives have placed growing weight on eliminating performance differences across protected groups (e.g., White House 2023). In our setting, however, requiring equal accuracy across patient types is not innocuous: it reshapes both the firm's investment incentives and physicians' deployment incentives, with direct consequences for patient welfare. This section therefore asks a simple question: holding fixed the liability framework, what are the welfare effects of imposing an equal-accuracy requirement?

We consider two policy routes to equal accuracy. The first is a direct mandate that forces equal accuracy in cases in which the firm would otherwise choose a disparate design. The second is an indirect approach that raises physician liability so that the firm endogenously prefers an equal-accuracy design. Because liability does not affect the firm's accuracy choice conditional on an equal-accuracy design, these two routes deliver the same equilibrium accuracy and thus the same expected patient welfare. We therefore focus, without loss of generality, on the direct mandate and characterize how imposing equal accuracy changes equilibrium accuracy and welfare.

Let $W_x(\rho_x^*)$ and $W_y(\rho_y^*)$ denote expected welfare for type-x and type-y patients, respectively, under the disparate design evaluated at the firm's optimal accuracies (ρ_x^*, ρ_y^*). Let ρ^m denote the accuracy level chosen under an equal-accuracy requirement, even in cases in which the firm would otherwise prefer a disparate design. Define $W_x(\rho^m)$ and $W_y(\rho^m)$ as the corresponding welfare levels under the equal-accuracy design. Closed-form expressions

for $W_x(\rho_x^*)$, $W_y(\rho_y^*)$, $W_x(\rho^m)$, and $W_y(\rho^m)$ are provided in the proof of Proposition 4. For ease of exposition, we define a threshold $\bar{f}$ by the indifference condition $W_y(\rho^m) = W_y(\rho_y^*)$ evaluated at $f = \bar{f}$.

Proposition 4. *When equal accuracy is mandated:*

(a) The optimal AI accuracy satisfies $\rho_y^* \leq \rho^m \leq \rho_x^*$.

(b) The expected patient surplus remains unchanged for both type-x and type-y patients when the physician is fairly certain about the appropriate treatment (that is, when α *is close to 0 or 1). It worsens (improves) for type-x (type-y) patients when the physician is highly uncertain (that is, when* α *is close to* $1/2$*). Otherwise, that is, when the uncertainty is moderate, type-x (type-y) patients are better (worse) off.*[8]

(c) Suppose $c < \frac{\kappa_y((2+\beta)\kappa_x+\kappa_y)\theta r}{\kappa_y(2\kappa_x+\kappa_y)-\beta\kappa_x^2}$ *and* $b > \frac{4c\kappa_y(\kappa_x+\kappa_y)}{\kappa_x(\kappa_y+\beta(\kappa_x+2\kappa_y))}$. *When* ℓ *is small, mandating equal accuracy yields three regimes. (1) If* $\frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2} < f < \frac{2c\kappa_x(\kappa_x+\kappa_y)}{b((2+\beta)\kappa_x+\kappa_y)}$, *then* $W_x(\rho^m) > W_x(\rho_x^*)$ *and* $W_y(\rho^m) < W_y(\rho_y^*)$. *(2) If* $\frac{2c\kappa_x(\kappa_x+\kappa_y)}{b((2+\beta)\kappa_x+\kappa_y)} < f < \bar{f}$, *then* $W_x(\rho^m) < W_x(\rho_x^*)$ *and* $W_y(\rho^m) < W_y(\rho_y^*)$; *at* $f = \frac{2c\kappa_x(\kappa_x+\kappa_y)}{b((2+\beta)\kappa_x+\kappa_y)}$, *type-x welfare is unchanged, and at* $f = \bar{f}$, *type-y welfare is unchanged. (3) Finally, if* $\bar{f} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y}{2\beta(b+\theta\ell)}\}$, *then* $W_x(\rho^m) < W_x(\rho_x^*)$ *and* $W_y(\rho^m) > W_y(\rho_y^*)$.[9]

Proposition 4(a) shows imposing an equal-accuracy requirement reshuffles performance across groups: relative to the firm's preferred disparate design, accuracy falls for type-x patients (so $\rho^m \leq \rho_x^*$) and rises for type-y patients (so $\rho^m \geq \rho_y^*$). A natural first reaction is therefore "type-y gains, type-x loses." Proposition 4(b) cautions against that conclusion. Whether a given patient benefits depends not only on accuracy but also on how the requirement changes the physician's use of AI along the margin.

The mechanism works through utilization. By raising accuracy for type-y patients and lowering it for type-x patients, the equal-accuracy requirement expands AI use for some type-y patients and contracts AI use for some type-x patients. For the newly treated type-y margin, welfare can fall when physician uncertainty is moderate: the incremental health improvement from relying on AI is then small, while the fixed cost of using AI, c, is paid

[8] The expressions for thresholds (which are different for type-x and type-y patients) that specify the range of α in which patient welfare improves, worsens, and remains unchanged are provided in the proof of Proposition 4.

[9] Comparisons of patient welfare over a wide range of (β, ℓ, f) values are provided in Section OA1 of the Online Appendix.

whenever the physician deploys the tool. For the dropped type-x margin, welfare can rise for the symmetric reason: these patients forgo a modest accuracy benefit but avoid incurring the cost c. In short, an equal-accuracy requirement can move patients onto (or off) AI precisely where the accuracy gains are too small (or too large) relative to the use cost.

Having established that welfare can move in either direction at the utilization margin, we turn to aggregate welfare by patient type. Proposition 4(c) shows that when ℓ is small, an equal-accuracy requirement can reduce aggregate welfare for both groups when the per-use payment f lies in an intermediate range. Two forces drive this outcome. First, the accuracy adjustment required to equalize performance is inherently asymmetric: because improving type-y accuracy is more expensive, parity is achieved primarily by cutting type-x accuracy, so $\rho_x^* - \rho^m > \rho^m - \rho_y^*$. Second, the requirement shifts utilization incentives. For type-y patients, higher accuracy raises welfare directly but also expands AI use, which can push deployment beyond the clinical margin; both effects strengthen with f, and the overuse channel can dominate when f is moderate. For type-x patients, reduced use can mitigate overuse, but the loss from lower accuracy becomes increasingly important as f induces larger cuts. When f is neither too small to matter nor large enough to generate substantial type-y accuracy gains, the combination of a sizeable decline in type-x accuracy and expanded (and potentially excessive) use for type-y patients can leave both groups worse off.

When f is sufficiently large, the balance shifts. The equal-accuracy requirement then induces a substantial reallocation of accuracy from type-x to type-y, yielding the intuitive aggregate ordering: $W_x(\rho^m) < W_x(\rho_x^*)$ while $W_y(\rho^m) > W_y(\rho_y^*)$. At the same time, as Proposition 4(b) emphasizes, these aggregate comparisons can mask heterogeneity: some type-y patients can still be worse off under the requirement because expanded reliance on AI occurs precisely where the clinical gains are small relative to the use cost.

Finally, when f is sufficiently small, welfare differences are driven less by accuracy and more by utilization. In this region, both gaps $\rho_x^* - \rho^m$ and $\rho^m - \rho_y^*$ are small, so the equal-accuracy requirement mainly shifts the frequency of use: it reduces AI use for type-x patients and increases it for type-y patients. Type-x patients can therefore benefit from fewer exposures to a low-accuracy tool on the margin, whereas type-y patients can be harmed because the increase in use is not accompanied by a commensurate improvement in accuracy. In this case, we may have $W_x(\rho^m) > W_x(\rho_x^*)$ yet $W_y(\rho^m) < W_y(\rho_y^*)$.

Next, we turn to a brief discussion of an alternative setting in which patient utility accounts for potential liability transfers. Under this formulation, the patient's utility and the

physician's payoff are updated according to Tables 4 and 5.

Table 4: The Patient's Expected Utility When the Physician Uses AI and Liability Transfers Are Included

	Physician's decision	
AI signal	T_1	T_2
$s_t = T_1$	$Q_{1\|1} \cdot b - c + Q_{2\|1} \cdot \mathbf{1}_{t=y}\,\ell$	$Q_{2\|1} \cdot b - c$
$s_t = T_2$	$Q_{1\|2} \cdot b - c$	$Q_{2\|2} \cdot b - c + Q_{1\|2} \cdot \mathbf{1}_{t=y}\,\ell$

Table 5: The Physician's Expected Payoff When Treating an Individual Patient Using AI and Liability Transfers Are Included

	Physician's decision	
AI signal	T_1	T_2
$s_t = T_1$	$Q_{1\|1} \cdot b - c + \theta r - (\theta - 1) Q_{2\|1} \cdot \mathbf{1}_{t=y}\,\ell$	$Q_{2\|1} \cdot b - c + \theta r$
$s_t = T_2$	$Q_{1\|2} \cdot b - c + \theta r$	$Q_{2\|2} \cdot b - c + \theta r - (\theta - 1) Q_{1\|2} \cdot \mathbf{1}_{t=y}\,\ell$

Note that if $\theta < 1$, the physician may use AI solely to exploit the patient's financial gain from the liability transfer, rather than to improve clinical outcomes. To rule out this uninteresting regime, we impose the parametric restriction $\theta > 1$. Under this condition, the physician's payoff matrix in Table 5 is structurally isomorphic to that in Table 3, with the weight on the liability term updated from θ to $\theta - 1$. Consequently, Lemma 2 continues to hold directly by updating the liability-adjusted selfishness parameter to $\theta - 1$, and the subsequent analytical insights follow analogously.

6. Model Extensions

In this section, we present three extensions to our base model. In the first extension, the AI firm also chooses the per-use price in addition to the two accuracy levels. In the second extension, we explore the liability level that maximizes aggregate patient welfare. Finally, we

consider patient-type-dependent priors for the physician. Proofs of the results in this section are in the Online Appendix (Sections OA2, OA3, and OA4).

6.1 AI Firm's Pricing Decision

This section endogenizes the per-use fee f in an extension in which patients pay for AI directly, so $c = f$. This structure captures emerging usage-based arrangements in which providers charge an out-of-pocket fee for AI-enhanced services. For example, RadNet (a large diagnostic imaging company) launched its Enhanced Breast Cancer Detection program in 2023 at a \$60 out-of-pocket fee for an AI mammography read and has since reduced the fee to \$40 (Cheatham 2024). The timing parallels the baseline model: the firm first chooses accuracy levels; it then sets the fee f; the physician observes (ρ_x, ρ_y, f) when deciding whether to use AI for a given patient. All other assumptions are unchanged. We summarize the core implications here and defer the technical derivations and additional comparative-statics details to Section OA2 of the Online Appendix.

Proposition 5. *Let $\ell^\dagger$ denote the design-switch cutoff at which the firm is indifferent between supplying a disparate design and an equal-accuracy design. Define*

$$D(\ell) := 2b^2\Big(\beta^2\kappa_x + \kappa_y\Big) - 4b\kappa_x\Big((\beta+1)\kappa_y - \beta^2\theta\ell\Big) + 2\beta^2\kappa_x\theta^2\ell^2$$

and the common factors $\mathcal{T} := \theta\,(\beta\ell - 2(\beta+1)r)\,/\,D(\ell)$ and $\mathcal{S} := \theta r\,/\,(2(\kappa_x+\kappa_y) - b(\beta+1))$. There exist parameter values such that the equilibrium accuracies are interior, $1/2 < \rho^*, \rho_x^*, \rho_y^* < 1$, and the following holds:

(a) If $\ell < \ell^\dagger$, the firm chooses a disparate design with

$$\rho_x^* = \frac{1}{2} + b\kappa_y\mathcal{T}, \qquad \rho_y^* = \frac{1}{2} + \beta\kappa_x(b+\theta\ell)\,\mathcal{T}, \qquad f^{\mathrm{D}} = b\kappa_x\kappa_y\mathcal{T}.$$

Moreover, $\rho_x^* > \rho_y^*$, and there are parameter values for which both f^{D} and ρ_x^* are non-monotone in ℓ.

(b) If $\ell \geq \ell^\dagger$, the firm supplies an equal-accuracy design with

$$\rho^* = \frac{1}{2} + (\beta+1)\,\mathcal{S} \qquad \textit{and} \qquad f^{\mathrm{E}} = (\kappa_x+\kappa_y)\,\mathcal{S}.$$

A key implication of Proposition 5 is that endogenizing the per-use fee can make both price and utilization respond non-monotonically to liability. When ℓ rises from a low level, expected utilization among type-y patients falls, weakening marginal demand; the firm then finds it profitable to cut the fee and to reallocate investment toward improving ρ_y^* to sustain type-y uptake. The lower fee, in turn, expands type-x usage, which allows the firm to economize on ρ_x^* while maintaining adoption. As ℓ becomes larger, however, the return to improving type-y accuracy strengthens and type-y utilization rebounds; the firm can then raise the fee, which dampens type-x demand and makes it attractive to increase ρ_x^* to restore type-x uptake.

We also find that, within the disparate-design regime, type-y utilization is non-monotone in ℓ for some parameter values: it decreases when ℓ is below a threshold, but increases once ℓ exceeds that threshold. Thus, the non-monotonicity in AI use for type-y patients persists even when the firm endogenizes the per-use fee f and patients bear that fee, so that $c = f$. As in the main model, an equal-accuracy requirement can still reduce welfare for both patient types when the AI firm endogenizes the per-use price.

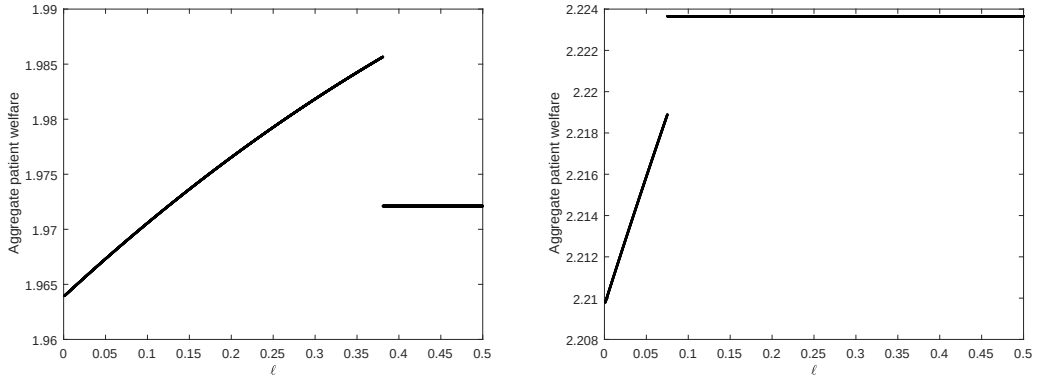
Next, we highlight two additional insights from the model with endogenous pricing. First, when the AI firm endogenously chooses its per-use price f, the equilibrium in which only type-x patients are served, which arises in the baseline model when ℓ is large and β is moderate, no longer occurs. Instead, both patient types are served, and an equal-accuracy AI outcome can be sustained. As a result, both the firm's revenue and AI equity improve. The additional pricing instrument allows the firm to profitably serve some type-y patients while preserving revenue from type-x patients through the joint choice of price and accuracy. Moreover, equal-accuracy AI becomes easier to sustain profitably because it mitigates the physician's liability concerns, encourages greater AI use, and allows the firm to further expand utilization through pricing.

Second, when f is endogenous, the welfare of type-x patients is also affected by liability. In particular, when liability is low, mandating equal accuracy can reduce type-x patients' welfare relative to the case with exogenous f. The reason is that, when liability is low, the equilibrium accuracy for type-x patients can be high; see Proposition 5. In that case, an equal-accuracy requirement may substantially reduce type-x accuracy, thereby lowering type-x welfare.

6.2 Aggregate-Welfare-Maximizing Liability

This subsection studies how the liability level ℓ should be set to maximize aggregate expected welfare. Two features of the model make the problem unusually sharp. First, holding the firm's AI design fixed, expected total welfare for type-x patients under a disparate design does not depend on ℓ, and aggregate welfare under an equal-accuracy design does not depend on ℓ. Second, Proposition 1 implies that increasing ℓ can change the firm's preferred design, inducing a switch from a disparate design to an equal-accuracy design at a threshold $\widetilde{\ell}$. As a result, liability affects aggregate welfare primarily through whether it triggers this design transition. We present the main economic intuition below and refer readers to Section OA3 of the Online Appendix for computational details and supplementary parameterizations.

Across the parameterizations we examine numerically, the welfare-maximizing liability lies at the boundary of this switch. Specifically, it is either the largest ℓ for which the firm still prefers a disparate design (see Figure 5a) or the smallest ℓ for which the firm prefers an equal-accuracy design (see Figure 5b). Thus, from the perspective of aggregate welfare, choosing ℓ amounts to choosing the design regime—keeping the system in the disparate-design region or inducing the shift to equal accuracy. Throughout, aggregate welfare refers to total patient health surplus, summed across both patient groups. We do not interpret it as full social welfare: doing so would require a common criterion encompassing patient health surplus, physician utility, firm profit, and any deadweight legal or compliance costs. We regard a full social-welfare analysis as a natural direction for future work.



(a) Case in which welfare is higher under a disparate design.

(b) Case in which welfare is higher under an equal-accuracy design.

Figure 5: Aggregate welfare as a function of liability ℓ. The upward-sloping segment is welfare under the firm's preferred disparate design; the flat segment is welfare once the firm switches to an equal-accuracy design, which does not vary with ℓ.

6.3 Type-Specific Priors

In this section, we allow physicians to hold type-specific priors about whether treatment T_1 is appropriate. This extension addresses the concern that imposing $\alpha_x = \alpha_y$ may understate human-driven disparities in baseline clinical beliefs. Specifically, we assume $\alpha_x \sim U(0, \overline{\alpha})$ and $\alpha_y \sim U(\underline{\alpha}, 1)$, where $\underline{\alpha} < 1/2 < \overline{\alpha}$. All other assumptions remain as in the main model. In particular, we maintain parameter restrictions under which, at every optimum characterized in this extension (and, under the disparate design, at every liability level up to the design-switch cutoff), demand for each patient type and the firm's profit are strictly positive, and the optimal accuracies are interior in $(1/2, 1)$. The numerical configurations reported in this extension satisfy these restrictions. Throughout this extension, we define $\tilde{f} := f/\overline{\alpha}$ and $\tilde{\beta} := \beta\overline{\alpha}/(1-\underline{\alpha})$ for the density-adjusted fee and composition. Because $\beta \in (0, 1)$ as in the main model, we assume $\tilde{\beta} < 1$ and $\overline{\alpha} + \underline{\alpha} > 1$.

We summarize our key findings and relegate our technical details to Section OA4 of the Online Appendix. The physician's optimal AI-use rule retains the same threshold form as in the baseline: AI is used only for an interior range of priors, and conditional on use, the physician follows the AI signal. Thus, downstream behavior is structurally unchanged, although the relevant thresholds shift with the group-specific prior distributions. Aggregating over these priors changes equilibrium AI-use volumes, but the firm's design problem remains qualitatively the same. The AI firm still chooses between a disparate design ($\rho_x > \rho_y$) and an equal-accuracy design ($\rho_x = \rho_y$), and there exists a liability cutoff $\widetilde{\ell}$ such that the firm prefers a disparate design for $\ell < \widetilde{\ell}$ and an equal-accuracy design otherwise. Comparative statics with respect to liability and reimbursement therefore continue to follow the baseline logic over the parameter regions we characterize in the Online Appendix: heterogeneous priors change the size of the adoption region rather than the underlying mechanism.

The welfare analysis carries over, with the integration limits adjusted to reflect the new prior supports, though the incidence of the welfare loss shifts. The main qualitative insights remain robust: at low levels, liability can reduce AI use for type-y patients, whereas at higher levels it can increase use through stronger upstream accuracy incentives; and an equal-accuracy mandate can still induce overuse and reduce welfare. In short, allowing $\alpha_x \neq \alpha_y$ preserves the use-distortion mechanism, and an equal-accuracy mandate can again harm both patient groups over a substantial region of the parameter space.

7. Concluding Remarks

Policymakers increasingly seek to address uneven algorithm performance across patient groups, yet the primary levers available in practice often operate at deployment: hospitals and physicians remain accountable for outcomes even when they rely on clinical decision-support tools. This paper studies the equilibrium consequences of such deployment-facing accountability by linking downstream physician reliance to upstream firm design. In a model with an AI firm and a physician, the Section 1557 clinical algorithm provision is captured as an asymmetric liability exposure that is triggered when reliance on a tool with unequal performance leads to inappropriate treatment for disadvantaged patients. Regulating use can reshape design, and regulating design can reshape use; welfare depends on the joint equilibrium.

Liability intended to protect disadvantaged patients can instead reduce their access to AI. By increasing the expected cost of relying on AI for type-y patients, liability can induce physicians to use AI less for the very group the policy aims to safeguard, even when the same tool would be deployed for type-x patients. The relationship between liability and use is also non-monotone. As liability rises, it can initially suppress reliance on AI for disadvantaged patients, but beyond a threshold it can induce the firm to reallocate investment toward their accuracy, improving performance and expanding adoption. A single instrument can therefore produce underuse at low levels of liability and renewed overuse at high ones, with a corrective range in between, depending on how strongly it feeds back to design incentives.

An equal-accuracy requirement is not an appropriate-use guarantee. Requiring equal accuracy across patient types necessarily reallocates model quality across groups; because improving type-y accuracy is more costly, parity is often achieved largely by reducing type-x accuracy. Equalizing measured performance also relaxes the liability channel that previously discouraged reliance for disadvantaged patients, while reimbursement continues to reward use. The result is a utilization response that can move patients onto (or off) AI precisely where clinical gains are small relative to the fixed cost of use. The mandate can therefore reduce aggregate welfare for both groups over parameter regions we characterize analytically: type-x patients lose from a substantial accuracy decline, whereas type-y patients can be harmed by expanded reliance on a still-imperfect tool.

When deployment incentives are shaped by reimbursement and accountability, standards aimed at algorithm performance need to be paired with instruments that govern use. In our

setting, liability standards that encourage firms to invest in disadvantaged-group accuracy need not align physicians' deployment choices with clinical value. Several complementary levers are available: reimbursement rules that attenuate incentives for indiscriminate use, auditing and monitoring requirements aimed at clinically meaningful endpoints, and other accountability mechanisms that discipline overuse. Without them, better measured performance need not translate into better patient welfare.

Although the motivating application is healthcare, the mechanism is broader: whenever professionals remain accountable for decisions informed by algorithms, deployment-facing accountability can generate feedback from use to design and back again. Analogous forces arise when judges rely on risk scores, when lenders and managers use credit and screening systems, and when employers deploy hiring tools. In each case, performance differences across groups can trigger legal or reputational exposure and thereby distort equilibrium reliance. Future work could extend the analysis to richer organizational environments (for example, hospitals with multiple clinicians), alternative contracting and pricing arrangements, and empirical measurement of the predicted non-monotone responses of both adoption and design to liability exposure.

Our analysis also points to several extensions that are especially relevant for AI. First, once an AI system has been trained, its marginal cost of use is conceivably low, whereas obtaining a second opinion from a human expert can entail a nontrivial cost. Allowing decision makers to choose between AI assistance and human peer review would introduce an additional margin of substitution and raise a distinct set of questions that we leave for future research. Second, clinical AI systems can be retrained as new data arrive. If liability discourages use for disadvantaged patients, those patients may become underrepresented in future training data, reducing subsequent accuracy and reinforcing the disparity the policy seeks to mitigate. Third, the AI signal may be more accurate than the physician's independent assessment for many patients. When physicians can adjust their own diagnostic effort, greater reliance on AI may reduce independent clinical scrutiny and create a new moral-hazard problem. Future work could also examine institutional arrangements that jointly determine reimbursement, pricing, utilization costs, and liability; allocate responsibility across physicians, hospitals, and developers through shared liability or ex ante subsidies; and link reputational returns to realized performance over time. A full social-welfare analysis is a further direction: treating our health-surplus measure as social welfare would require a common criterion spanning patient surplus, physician utility, firm profit, and deadweight legal and compliance costs.

Funding and Competing Interests

All authors certify that they have no affiliations with or involvement in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript. The authors have no funding to report.

Appendix

PROOF OF LEMMA 1. Using the physician's expected payoff when not using AI (from eq. (3)) and when using AI (from eq. (4)), we compute the difference $U_x - U$ which is given by

$$U_x - U = \begin{cases} \theta r - c, & \text{if } \alpha \leq 1 - \rho_x \\ (\rho_x + \alpha - 1) \cdot b - c + \theta r, & \text{if } 1 - \rho_x < \alpha \leq 1/2 \\ (\rho_x - \alpha) \cdot b - c + \theta r, & \text{if } 1/2 < \alpha \leq \rho_x \\ \theta r - c, & \text{otherwise.} \end{cases} \tag{A1}$$

Note that $\theta r - c < 0$; therefore, the physician does not use AI for $\alpha \leq 1 - \rho_x$ and for $\alpha > \rho_x$. In the second case of eq. (A1), that is, if $1 - \rho_x < \alpha \leq 1/2$, the physician does not use AI when $\alpha \leq \frac{(1-\rho_x)b+c-\theta r}{b}$ and uses AI when $\alpha \geq \frac{(1-\rho_x)b+c-\theta r}{b}$. In the third case of eq. (A1), that is, if $1/2 < \alpha \leq \rho_x$, the physician does not use AI when $\alpha > \frac{\rho_x b-(c-\theta r)}{b}$ and uses AI when $\alpha \leq \frac{\rho_x b-(c-\theta r)}{b}$. Therefore, we have

- When $c - \theta r < (\rho_x - 1/2)b$, we have $\frac{(1-\rho_x)b+c-\theta r}{b} < 1/2$ and $\frac{\rho_x b-(c-\theta r)}{b} > 1/2$. Then, the physician does not use AI if $\alpha \leq \frac{(1-\rho_x)b+c-\theta r}{b}$, uses AI if $\frac{(1-\rho_x)b+c-\theta r}{b} < \alpha \leq \frac{\rho_x b-(c-\theta r)}{b}$, and does not use AI again if $\alpha > \frac{\rho_x b-(c-\theta r)}{b}$.
- When $c - \theta r \geq (\rho_x - 1/2)b$, we have $\frac{(1-\rho_x)b+c-\theta r}{b} \geq 1/2$ and $\frac{\rho_x b-(c-\theta r)}{b} \leq 1/2$. Then, the physician does not use AI for all patients $\alpha \in [0, 1]$.

Under the assumption that the physician uses AI for at least some type-y patients (that is, $d_y^{\mathrm{D}} > 0$, equivalently, $c - \theta r < (\rho_y - \frac{1}{2})b - (1 - \rho_y)\theta\ell$) and given $\rho_x > \rho_y$, we have $c - \theta r < (\rho_x - 1/2)b$. Therefore, we derive the physician's AI use for type-x patients in this lemma.

Next, we check whether the physician will follow or reject the AI signal when using AI. In the second and third cases of eq. (A1), the physician follows the AI signal regardless of the signal. These are also the only scenarios in which the physician uses AI. Therefore, the physician follows the AI signal for type-x patients, whenever using AI. *Q.E.D.*

PROOF OF LEMMA 2. Using the physician's expected payoff when not using AI (from eq. (3)) and when using AI (from eq. (5)), we compute the difference $U_y - U$ which is given by:

$$U_y - U = \begin{cases} \theta r - c - \alpha(1 - \rho_y)\theta\ell, & \text{if } \alpha \leq \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} \\ (\rho_y + \alpha - 1) \cdot b - c + \theta r - (1 - \rho_y)\theta\ell, & \text{if } \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} < \alpha \leq 1/2 \\ (\rho_y - \alpha) \cdot b - c + \theta r - (1 - \rho_y)\theta\ell, & \text{if } 1/2 < \alpha \leq \frac{b\rho_y}{b+(1-\rho_y)\theta\ell} \\ \theta r - c - (1 - \alpha)(1 - \rho_y)\theta\ell, & \text{otherwise.} \end{cases} \tag{A2}$$

Note that $\theta r - c - \alpha(1-\rho_y)\theta\ell < 0$; therefore, the physician does not use AI for $\alpha \leq \frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell}$. In addition, $\theta r - c - (1-\alpha)(1-\rho_y)\theta\ell < 0$, which implies the physician does not use AI for $\alpha > \frac{b\rho_y}{b+(1-\rho_y)\theta\ell}$.

In the second case of eq. (A2), that is, if $\frac{(1-\rho_y)(b+\theta\ell)}{b+(1-\rho_y)\theta\ell} < \alpha \leq 1/2$, the physician does not use AI when $\alpha \leq \frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b}$ and uses AI when $\alpha \geq \frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b}$. In the third case of eq. (A2), that is, if $1/2 < \alpha \leq \frac{b\rho_y}{b+(1-\rho_y)\theta\ell}$, the physician does not use AI when $\alpha > \frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b}$ and uses AI when $\alpha \leq \frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b}$. Therefore, we have

- When $c-\theta r < (\rho_y - \frac{1}{2})b - (1-\rho_y)\theta\ell$, we have $\frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b} < 1/2 < \frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b}$. Then, the physician does not use AI if $\alpha \leq \frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b}$, uses AI if $\frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b} < \alpha < \frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b}$, and does not use AI again if $\alpha > \frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b}$.
- When $c-\theta r > (\rho_y - \frac{1}{2})b - (1-\rho_y)\theta\ell$, we have $\frac{(1-\rho_y)(b+\theta\ell)+c-\theta r}{b} > 1/2$ and $\frac{\rho_y b-(1-\rho_y)\theta\ell-c+\theta r}{b} < 1/2$. Then, the physician does not use AI for all patients $\alpha \in [0,1]$.

Under the assumption that the physician uses AI for at least some type-y patients (that is, $d_y^{\mathrm{D}} > 0$, equivalently, $c-\theta r < (\rho_y - \frac{1}{2})b - (1-\rho_y)\theta\ell$), we derive the physician's AI use for type-y patients in this lemma.

Next, we check whether the physician will follow or reject the AI signal when using AI. In the second and third cases of eq. (A2), the physician follows the AI signal regardless of the signal. These are also the only scenarios in which the physician might choose to use AI. Therefore, the physician follows the AI signal for type-y patients whenever using AI. *Q.E.D.*

Proof of Proposition 1. (a) Suppose the AI firm supplies a disparate algorithm. It sets ρ_x and ρ_y to maximize the profit in eq. (1), where d_x^{D} and d_y^{D} are given in eq. (6). The objective is strictly concave in ρ_x and ρ_y, so the first-order conditions are sufficient. Setting $\partial\pi^{\mathrm{D}}/\partial\rho_x = 0$ and $\partial\pi^{\mathrm{D}}/\partial\rho_y = 0$ yields $\rho_x^* = \frac{1}{2} + \frac{f}{\kappa_x}$ and $\rho_y^* = \frac{1}{2} + \frac{\beta f(b+\theta\ell)}{b\kappa_y}$. The interior conditions $\rho_x^* < 1$ and $\rho_y^* < 1$ are equivalent to $f < \kappa_x/2$ and $f < \frac{b\kappa_y}{2\beta(b+\theta\ell)}$, respectively.

Next, to ensure that the firm earns a positive profit from serving type-y patients at ρ_y^*, that is, $\pi_y(\rho_y^*) > 0$, it suffices that $f > \frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2}$. This condition also guarantees that the physician uses AI for at least some type-y patients under the induced accuracy choice (equivalently, $d_y^{\mathrm{D}}|_{\rho_y=\rho_y^*} > 0$). Otherwise, the optimal choice sets ρ_y arbitrarily close to $\frac{1}{2}$.

We first note that no strictly reverse-disparate design ($\rho_y > \rho_x$) can be optimal. In that region the disparity-triggered liability channel is inactive, so the objective is the no-liability profit, which is strictly concave with unconstrained maximizers $\rho_x^0 = \frac{1}{2} + \frac{f}{\kappa_x}$ and $\rho_y^0 = \frac{1}{2} + \frac{\beta f}{\kappa_y}$. Because $\kappa_y > \kappa_x$ and $\beta \leq 1$, we have $\frac{f}{\kappa_x} > \frac{\beta f}{\kappa_y}$, so $\rho_x^0 > \rho_y^0$ and

the unconstrained maximizer lies outside this region. By strict concavity the constrained maximizer lies on the boundary $\rho_x = \rho_y$, which is the equal-accuracy design. It therefore suffices to compare the disparate and equal-accuracy designs. We then show that whenever a disparate design is optimal, it must satisfy $\rho_x^* > \rho_y^*$. If $\rho_x^* = \rho_y^*$ under the disparate-design best response, then the firm can instead supply the same accuracy level as an equal-accuracy design, which eliminates the liability channel and weakly increases physician demand; hence $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) < \pi^{\mathrm{E}}(\rho_x^*) \leq \max_\rho \pi^{\mathrm{E}}(\rho) = \pi^{\mathrm{E}}(\rho^*)$. This contradicts optimality of a disparate design. To preclude the reversal $\rho_x^* < \rho_y^*$, we invoke a continuity argument. At the boundary $\ell = 0$, the optimal accuracies yield a positive gap, $\rho_x^* - \rho_y^* = f(1/\kappa_x - \beta/\kappa_y) > 0$. Because $\rho_x^* - \rho_y^*$ is continuous and strictly decreasing in ℓ, the gap cannot change sign and become negative without passing through zero. Since exact equality $\rho_x^* = \rho_y^*$ is strictly ruled out by our previous contradiction argument, it follows by the Intermediate Value Theorem that $\rho_x^* > \rho_y^*$ must hold strictly throughout the entire feasible domain. Therefore, if the disparate design is optimal, we must have $\rho_x^* > \rho_y^*$.

(b) Now suppose the firm supplies an equal-accuracy algorithm, so $\rho_x = \rho_y \equiv \rho$. Maximizing eq. (2) and setting $\partial \pi^{\mathrm{E}}(\rho)/\partial \rho = 0$ yields $\rho^* = \frac{1}{2} + \frac{(1+\beta)f}{\kappa_x + \kappa_y}$. The interior condition $\rho^* < 1$ is equivalent to $f < \frac{\kappa_x + \kappa_y}{2(1+\beta)}$. Note that $\frac{\kappa_x + \kappa_y}{2(1+\beta)} > \frac{\kappa_x}{2}$ because $\kappa_y > \beta \kappa_x$; hence the standing condition $f < \kappa_x/2$ (which ensures $\rho_x^* < 1$) already implies $f < \frac{\kappa_x + \kappa_y}{2(1+\beta)}$, and the interior constraint for the equal-accuracy design can be omitted.

Next, we show that there exists a cutoff $\widetilde{\ell}$ such that a disparate design is optimal for $\ell < \widetilde{\ell}$ and an equal-accuracy design is optimal for $\ell \geq \widetilde{\ell}$. As $\ell \to 0$, the disparate-design profit converges to the equal-accuracy objective evaluated at potentially different accuracies across types, that is, $\pi^{\mathrm{D}}(\rho_x, \rho_y)|_{\ell \to 0} \to \pi^{\mathrm{E}}(\rho_x, \rho_y)$, while the equal-accuracy design restricts the firm to $\rho_x = \rho_y$. Because the feasible set under a disparate design weakly contains that under equal accuracy, we have $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*)|_{\ell \to 0} \geq \pi^{\mathrm{E}}(\rho^*)$, with strict inequality whenever the equality constraint $\rho_x = \rho_y$ binds.

Next, under a disparate design, the firm's optimal profit is decreasing in ℓ. The type-x component is independent of ℓ, so it suffices to study the type-y component. By the envelope theorem,

$$\frac{d}{d\ell} \pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) = \left. \frac{\partial \pi^{\mathrm{D}}(\rho_x, \rho_y)}{\partial \ell} \right|_{\rho_x = \rho_x^*,\, \rho_y = \rho_y^*} = -\frac{2\theta f \beta}{b} (1 - \rho_y^*) < 0,$$

where the negative sign strictly holds in the interior region in which $\rho_y^* < 1$. By contrast, $\pi^{\mathrm{E}}(\rho^*)$ does not depend on ℓ. Because $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)$ is strictly decreasing in ℓ under

this interior condition, there is at most one cutoff $\widetilde{\ell}$ at which $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) = \pi^{\mathrm{E}}(\rho^*)$, and the profit ranking switches at that point. Outside this interior region (where $\rho_y^* \geq 1$), although $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)$ becomes a convex parabola in ℓ that could theoretically admit a second crossing, such a region is structurally infeasible and can be disregarded, thereby establishing the unique design cutoff.

We conclude the proof by establishing the existence of the cutoff $\widetilde{\ell}$ within the feasible domain. A sufficient condition for $\widetilde{\ell}$ to be interior is $\pi^{\mathrm{E}}(\rho^*) > \pi_x(\rho_x^*)$, which simplifies to $\beta(\beta+2)\kappa_x > \kappa_y$ and $f > \frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)}$. This parameter restriction ensures that the profit from type-y patients remains strictly positive ($\pi_y(\rho_y^*) > 0$) at the point of indifference where $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) = \pi^{\mathrm{E}}(\rho^*)$ (i.e., at $\ell = \widetilde{\ell}$). Consequently, both $\pi_y(\rho_y^*) > 0$ and $\rho_y^* < 1$ hold for all $\ell < \widetilde{\ell}$, where the latter inequality follows from our earlier result that $\rho_y^* < \rho_x^* < 1$ whenever a disparate design strictly dominates. Conversely, if this condition is violated, an interior cutoff $\widetilde{\ell}$ fails to exist. In that scenario, an equal-accuracy design never becomes optimal as ℓ increases; instead, the firm continuously deploys a disparate algorithm while completely abandoning the type-y market to serve type-x patients exclusively. *Q.E.D.*

Proof of Proposition 2. Under the disparate design, AI use for type-y patients is given by eq. (6) as $d_y^{\mathrm{D}} = \beta \cdot \frac{(2\rho_y^*-1)b-2(1-\rho_y^*)\theta\ell-2c+2\theta r}{b}$, where $\rho_y^* = \frac{1}{2} + \frac{\beta f(b+\theta\ell)}{b\kappa_y}$ from Proposition 1. Note d_y^{D} is convex in ℓ, implying that it has at most one turning point. Taking the first-order derivative of d_y^{D} with respect to ℓ yields $\frac{\partial d_y^{\mathrm{D}}}{\partial \ell} > 0$ if and only if $\ell > \bar{\ell} := \frac{b(\kappa_y-4\beta f)}{4\theta\beta f}$. We require $\bar{\ell} > 0$ to make sure that d_y^{D} has one unique turning point at $\ell = \bar{\ell}$, which gives us $f < \frac{\kappa_y}{4\beta}$. Thus, holding the AI firm's design fixed at the disparate regime, liability raises type-y AI use when ℓ exceeds $\bar{\ell}$ and lowers it when $\ell < \bar{\ell}$.

To establish the claimed pattern around $\bar{\ell}$ in equilibrium, note that the parameter restrictions in Proposition 1 imply: (i) $\frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y}{2\beta(b+\theta\ell)}\}$, (ii) the AI firm strictly prefers the disparate design at $\ell = \bar{\ell}$ (that is, $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ at $\ell = \bar{\ell}$), and (iii) $\beta(\beta+2)\kappa_x > \kappa_y$ and $f > \frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)}$.

We first establish conditions surrounding the restriction (i). We can substantiate that $f < \frac{b\kappa_y}{2\beta(b+\theta\ell)}|_{\ell=\bar{\ell}}$, which means $\rho_y^* < 1$ at $\ell = \bar{\ell}$ (by Proposition 1), which also means that $\rho_y^* < 1$ must hold for $\ell < \bar{\ell}$ (as ρ_y^* is increasing in ℓ). Furthermore, given $f > \frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2}$ is equivalent to $\pi_y(\rho_y^*) > 0$, we can omit the consideration of $f > \frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2}|_{\ell=\bar{\ell}}$. This is because $\pi_y(\rho_y^*)|_{\ell=\bar{\ell}} = [\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\bar{\ell}} + (\pi^{\mathrm{E}}(\rho^*) - \pi_x(\rho_x^*)) > 0$, where the inequality holds because the first term is greater than 0 (we will guarantee this by discussing

the restriction (ii) in what follows) and the second term is also greater than 0 by the restriction (iii) above. Moreover, $\frac{\partial \pi_y(\rho_y^*)}{\partial \ell} = \frac{2\beta f\theta}{b}\left(\rho_y^* - 1\right) < 0$ whenever $\rho_y^* < 1$, which holds on $[0, \bar{\ell}]$ as just established. Hence $\pi_y(\rho_y^*)|_\ell > \pi_y(\rho_y^*)|_{\ell=\bar{\ell}} > 0$ for every $\ell < \bar{\ell}$, so the participation constraint is slack throughout. Therefore, for the restriction (i), the only constraint is $f < \frac{\kappa_x}{2}$.

We then establish conditions surrounding the restriction (ii). We obtain $[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\bar{\ell}} = A \cdot f^2 + \beta f - \frac{3\kappa_y}{16}$, where $A = \frac{\kappa_y - \beta(\beta+2)\kappa_x}{\kappa_x(\kappa_x+\kappa_y)}$. Letting $[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\bar{\ell}} = 0$ yields $f = f_a$ or $f = f_b$, where $f_a := \frac{2\beta\kappa_x(\kappa_x+\kappa_y)}{4\beta(\beta+2)\kappa_x - 4\kappa_y} - \frac{\sqrt{\kappa_x(\kappa_x+\kappa_y)\left(\beta^2\kappa_x(4\kappa_x+\kappa_y) - 6\beta\kappa_x\kappa_y + 3\kappa_y^2\right)}}{4\beta(\beta+2)\kappa_x - 4\kappa_y}$ and $f_b := \frac{2\beta\kappa_x(\kappa_x+\kappa_y)}{4\beta(\beta+2)\kappa_x - 4\kappa_y} + \frac{\sqrt{\kappa_x(\kappa_x+\kappa_y)\left(\beta^2\kappa_x(4\kappa_x+\kappa_y) - 6\beta\kappa_x\kappa_y + 3\kappa_y^2\right)}}{4\beta(\beta+2)\kappa_x - 4\kappa_y}$. According to the restriction (iii), we have $A < 0$ and we can further obtain that $f_a < f_b$. Then $[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\bar{\ell}} > 0$ is equivalent to $f_a < f < f_b$. We can verify that $f_b > \min\{\frac{\kappa_x}{2}, \frac{\kappa_y}{4\beta}\}$. Thus, to ensure $[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\bar{\ell}} > 0$, we only need to impose an additional condition $f > f_a$. Therefore, under the aforementioned conditions, we can obtain $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*)|_{\ell<\bar{\ell}} > \pi^{\mathrm{D}}(\rho_x^*, \rho_y^*)|_{\ell=\bar{\ell}} > \pi^{\mathrm{E}}(\rho^*)$, which means that the AI firm strictly prefers the disparate design if $\ell < \bar{\ell}$.

Define $f_l := \max\left\{\frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x - \kappa_y)},\ f_a\right\}$. So far, we have proved that when $\beta(\beta+2)\kappa_x > \kappa_y$ and $f_l < f < \min\{\frac{\kappa_x}{2}, \frac{\kappa_y}{4\beta}\}$, AI use for disadvantaged patients is decreasing in ℓ for $\ell \le \bar{\ell}$. We next prove that AI use for disadvantaged patients is (weakly) increasing in ℓ for $\ell > \bar{\ell}$.

From the preceding analysis, when ℓ increases marginally beyond $\bar{\ell}$, the interior conditions $\pi_y(\rho_y^*) > 0$, $\rho_y^* < 1$, and $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ continue to hold. By continuity, there exists a neighborhood to the right of $\bar{\ell}$ where AI use among type-y patients is strictly increasing in ℓ. As ℓ increases further, one of these performance or participation constraints must eventually bind. From the proof of Proposition 1, $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ implies $\rho_x^* > \rho_y^*$. Given the interior requirement $\rho_x^* < 1$, it follows immediately that $\rho_y^* < 1$ is satisfied whenever the disparate design remains strictly optimal. Hence, either the regime-switching condition $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ or the market-viability condition $\pi_y(\rho_y^*) > 0$ must be the first to bind. Because parameter restriction (iii) ensures that $\pi^{\mathrm{E}}(\rho^*) > \pi_x(\rho_x^*)$, the profit under the equal-accuracy algorithm strictly dominates the profit from serving type-x patients alone. Consequently, $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ must be the first condition to fail. At this crossing point, the equal-accuracy algorithm becomes optimal, and AI use among type-y patients remains constant thereafter. Because AI use may exhibit a discrete upward jump at this regime switch (by the same argument as in the proof of part (a) of Proposition 4), AI use among type-y

patients is weakly increasing in ℓ for all $\ell > \bar{\ell}$. The proposition then follows[10]. *Q.E.D.*

Proof of Lemma 3. Using the altruistic physician's expected payoff when not using AI (from eq. (3)) and when using AI (from eq. (8)), we compute the difference $U^{\text{al}} - U$, given by

$$U^{\text{al}} - U = \begin{cases} -c, & \text{if } \alpha \leq 1 - \rho_t \\ (\rho_t + \alpha - 1) \cdot b - c, & \text{if } 1 - \rho_t < \alpha \leq 1/2 \\ (\rho_t - \alpha) \cdot b - c, & \text{if } 1/2 < \alpha \leq \rho_t \\ -c, & \text{otherwise.} \end{cases} \tag{A3}$$

It is straightforward that the altruistic physician does not use AI in the first and fourth cases, that is, when $\alpha \leq 1 - \rho_t$ or $\alpha > \rho_t$. In the second case of eq. (A3), the physician does not use AI when $\alpha \leq \frac{(1-\rho_t)b+c}{b}$ and uses AI otherwise. In the third case of eq. (A3), the physician does not use AI when $\alpha > \frac{\rho_t b - c}{b}$ and uses AI otherwise. Therefore, we have

- When $c < (\rho_t - 1/2) \cdot b$, we have $\frac{(1-\rho_t)b+c}{b} < 1/2$ and $\frac{\rho_t b - c}{b} > 1/2$. Then, the physician does not use AI if $\alpha \leq \frac{(1-\rho_t)b+c}{b}$, uses AI if $\frac{(1-\rho_t)b+c}{b} < \alpha \leq \frac{\rho_t b - c}{b}$, and does not use AI again if $\alpha > \frac{\rho_t b - c}{b}$.
- When $c \geq (\rho_t - 1/2) \cdot b$, we have $\frac{(1-\rho_t)b+c}{b} \geq 1/2$ and $\frac{\rho_t b - c}{b} \leq 1/2$. Then, the physician does not use AI for all patients $\alpha \in (0, 1)$.

Next, we check whether the physician will follow or reject the AI signal when using AI. In the second and third cases of eq. (A3), the physician follows the AI signal regardless of the signal. These are also the only scenarios in which the physician might choose to use AI. Therefore, the physician follows the AI signal for type-t patients whenever using AI. *Q.E.D.*

Proof of Proposition 3. (a) Under an equal-accuracy design, the disparity-triggered liability channel does not apply. An impurely altruistic physician therefore bases AI use for both patient types on the decision rule in Lemma 1. Comparing this rule with the corresponding benchmark for an altruistic physician in Lemma 3 implies that the impurely altruistic physician uses AI for a (weakly) larger set of patients, and hence overuses AI.

(b) Under a disparate design, comparing Lemma 2 with Lemma 3 shows that the impurely altruistic physician uses AI appropriately for type-y patients if and only if the reimbursement exactly offsets the expected liability cost on the margin, that is, $(1 - \rho_y^*)\ell = r$,

[10] A numerical instance illustrates the existence of this region. When $b = 2$, $\theta = 1$, $c = 0.60$, $r = 0.5$, $\kappa_x = 1.3$, $\kappa_y = 1.705$, $f = 0.6$, and $\beta = 0.68$, the physician uses AI for fewer type-y patients if $0 < \ell < 0.089$ and for more type-y patients if $0.089 < \ell < 0.375$. The equal-accuracy design becomes optimal if $\ell \geq 0.375$.

which is equivalent to $\frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y} = \frac{r}{\ell}$. By part (a), the physician overuses AI for type-x patients.[11]

(c) From the argument in part (b), under the disparate design the physician overuses AI for type-y patients when $\left(\frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y}\right)\ell < r$ and underuses AI when $\left(\frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y}\right)\ell > r$. The left-hand side is concave in ℓ, so the equality $\left(\frac{1}{2} - \frac{\beta f(b+\theta\ell)}{b\kappa_y}\right)\ell = r$ has (at most) two solutions; let ℓ_s denote the smaller root and ℓ_h denote the larger root,

$$\ell_s = \frac{b\kappa_y\left(1 - \sqrt{\frac{b(\kappa_y - 2\beta f)^2 - 16\beta f\kappa_y\theta r}{b\kappa_y^2}}\right)}{4\beta f\theta} - \frac{b}{2\theta} \text{ and } \ell_h = \frac{b\kappa_y\left(1 + \sqrt{\frac{b(\kappa_y - 2\beta f)^2 - 16\beta f\kappa_y\theta r}{b\kappa_y^2}}\right)}{4\beta f\theta} - \frac{b}{2\theta}.$$

In what follows, we prove that the physician first overuses AI when $\ell < \ell_s$ and then underuses AI when $\ell_s < \ell < \ell_m$ and overuses AI again when $\ell > \ell_m$, where ℓ_m is a threshold.

We start by establishing the claimed pattern around ℓ_s in equilibrium. Letting $\ell_s > 0$ yields $f < \frac{b\kappa_y + 4\kappa_y\theta r}{2b\beta} - \frac{\kappa_y\sqrt{2b\theta r + 4\theta^2 r^2}}{b\beta}$. In addition, note that the parameter restrictions in Proposition 1 imply: (i) $\frac{b\kappa_y(2c - 2\theta r + \theta\ell)}{\beta(b+\theta\ell)^2} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y}{2\beta(b+\theta\ell)}\}$, (ii) the AI firm strictly prefers the disparate design at $\ell = \ell_s$ (that is, $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ at $\ell = \ell_s$), and (iii) $\beta(\beta+2)\kappa_x > \kappa_y$ and $f > \frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x - \kappa_y)}$.

We first establish conditions surrounding the restriction (i). We can verify that $f < \frac{b\kappa_y}{2\beta(b+\theta\ell)}|_{\ell=\ell_s}$, which is equivalent to $\rho_y^*|_{\ell=\ell_s} = 1 - r/\ell_s < 1$; this also means that $\rho_y^* < 1$ must hold for $\ell < \ell_s$ (as ρ_y^* is increasing in ℓ). In addition, when $\ell = \ell_s$, considering $f > \frac{b\kappa_y(2c - 2\theta r + \theta\ell)}{\beta(b+\theta\ell)^2}$ or equivalently $\pi_y(\rho_y^*) > 0$ (by Proposition 1) yields

$$\frac{4\beta^2 f^2}{\kappa_y} + 4\beta f + (\kappa_y - 2\beta f)\sqrt{\frac{b(\kappa_y - 2\beta f)^2 - 16\beta f\kappa_y\theta r}{b\kappa_y^2}} > \kappa_y + \frac{8\beta f(2c - \theta r)}{b}; \tag{A4}$$

therefore, under the aforementioned condition, we can derive $\pi_y(\rho_y^*)|_{\ell<\ell_s} > \pi_y(\rho_y^*)|_{\ell=\ell_s} > 0$, which means that the firm earns a positive profit from serving type-y patients at ρ_y^* if $\ell < \ell_s$.

We then establish conditions surrounding the restriction (ii). Considering $[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^*)]|_{\ell=\ell_s}$ yields

$$\frac{8f^2}{\kappa_x} + \frac{4\beta f(\beta f + \kappa_y)}{\kappa_y} + (\kappa_y - 2\beta f)\sqrt{\frac{b(\kappa_y - 2\beta f)^2 - 16\beta f\kappa_y\theta r}{b\kappa_y^2}} > \kappa_y + \frac{8\beta f\theta r}{b} + \frac{8(1+\beta)^2 f^2}{\kappa_x + \kappa_y}. \tag{A5}$$

[11] A numerical example is given by $b = 1.8$, $\theta = 1$, $r = 0.1$, $\kappa_x = 2$, $\kappa_y = 3$, $f = 0.8$, $\beta = 0.6$, $c = 0.11$, and $\ell = 0.32$, under which AI is used appropriately for type-y patients.

Therefore, under the aforementioned condition, we can obtain

$$\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*)|_{\ell<\ell_s} > \pi^{\mathrm{D}}(\rho_x^*, \rho_y^*)|_{\ell=\ell_s} > \pi^{\mathrm{E}}(\rho^*),$$

which means that the AI firm strictly prefers the disparate design if $\ell < \ell_s$.

So far, we have the following conditions: (A4), (A5), and $f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y+4\kappa_y\theta r}{2b\beta} - \frac{\kappa_y\sqrt{2b\theta r+4\theta^2 r^2}}{b\beta}\}$, under which the physician overuses AI when $\ell < \ell_s$. We next prove that the physician underuses AI when $\ell_s < \ell < \ell_m$ and overuses AI again when $\ell > \ell_m$.

From the above proof, when ℓ increases marginally from ℓ_s, the conditions $\pi_y(\rho_y^*) > 0$, $\rho_y^* < 1$, and $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ continue to hold. By continuity, there therefore exists a neighborhood to the right of ℓ_s in which the physician underuses AI as ℓ increases from ℓ_s. As ℓ increases further, one of the preceding conditions must eventually fail. By an argument similar to that in the proof of Proposition 2, the restriction (iii) (i.e., $\beta(\beta+2)\kappa_x > \kappa_y$ and $f > \frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)}$) ensures that $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ is the first and only condition that can be violated as ℓ increases. Once this condition is violated, Part (a) implies that physician overuse of AI reappears. Depending on the relationship between the violation point and ℓ_h, two cases arise:

- Suppose $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ is violated before $\ell = \ell_h$. Then ℓ_m coincides with the violation point. In this case, the physician overuses AI for $\ell < \ell_s$, underuses AI for $\ell_s < \ell < \ell_m$, and overuses AI again for $\ell > \ell_m$. The first two regions arise under the disparate algorithm, whereas the final overuse region is induced by the equal-accuracy algorithm.

- Suppose $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) > \pi^{\mathrm{E}}(\rho^*)$ is violated after $\ell = \ell_h$. Then $\ell_m = \ell_h$. In this case, the physician overuses AI for $\ell < \ell_s$, underuses AI for $\ell_s < \ell < \ell_m$, and overuses AI again for $\ell > \ell_m$. The re-emergence of overuse initially occurs under the disparate algorithm and subsequently persists under the equal-accuracy algorithm after the algorithmic switch.

Combining the above two cases establishes the proposition under the conditions $\beta(\beta+2)\kappa_x > \kappa_y$, (A4), (A5), and $\frac{2\beta(c-\theta r)\kappa_x(\kappa_x+\kappa_y)}{b(\beta(\beta+2)\kappa_x-\kappa_y)} < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y+4\kappa_y\theta r}{2b\beta} - \frac{\kappa_y\sqrt{2b\theta r+4\theta^2 r^2}}{b\beta}\}$. This completes the proof[12]. *Q.E.D.*

[12]For example, when $b = 1.8$, $\theta = 1$, $r = 0.1$, $\kappa_x = 2$, $\kappa_y = 3$, $f = 0.8$, $\beta = 0.6$, and $c = 0.11$, the physician overuses AI when $\ell < 0.32$, underuses AI when $0.32 < \ell < 0.40$, and overuses AI again when $\ell \geq 0.40$, where the equal-accuracy design becomes optimal.

Proof of Proposition 4. Using expected utilities for different patient types in Table 2 and AI accuracy expressions in Proposition 1, we compute the expected patient welfare. By Lemmas 1 and 2, when the physician relies on an accuracy ρ without the disparity-liability channel, she uses AI exactly on $(\alpha^-(\rho), \alpha^+(\rho))$, where

$$\alpha^-(\rho) := \frac{(1-\rho)b + c - \theta r}{b}, \qquad \alpha^+(\rho) := \frac{\rho b - c + \theta r}{b};$$

under a disparate design she uses AI for type-y patients exactly on (α_y^-, α_y^+), where

$$\alpha_y^- := \frac{(1-\rho_y^*)(b+\theta\ell) + c - \theta r}{b}, \qquad \alpha_y^+ := \frac{\rho_y^* b - (1-\rho_y^*)\theta\ell - c + \theta r}{b}.$$

Accordingly,

$$\begin{aligned} W_x(\rho) &= \int_0^{\alpha^-(\rho)} \Big((1-\alpha)b\Big)d\alpha + \int_{\alpha^-(\rho)}^{\alpha^+(\rho)} \Big(\rho b - c\Big)d\alpha + \int_{\alpha^+(\rho)}^1 \Big(\alpha b\Big)d\alpha, \\ W_y(\rho_y^*) &= \beta \cdot \Big[\int_0^{\alpha_y^-} \Big((1-\alpha)b\Big)d\alpha + \int_{\alpha_y^-}^{\alpha_y^+} \Big(\rho_y^* b - c\Big)d\alpha + \int_{\alpha_y^+}^1 \Big(\alpha b\Big)d\alpha\Big]. \end{aligned} \tag{A6}$$

The four objects we compare are $W_x(\rho_x^*)$, $W_y(\rho_y^*)$, $W_x(\rho^m)$, and, because the mandate deactivates the disparity-liability channel for type-y patients, $W_y(\rho^m) = \beta\, W_x(\rho^m)$, where $\rho_x^* = \frac{1}{2} + \frac{f}{\kappa_x}$, $\rho_y^* = \frac{1}{2} + \frac{\beta f(b+\theta\ell)}{b\kappa_y}$ and $\rho^m = \frac{1}{2} + \frac{(1+\beta)f}{\kappa_x+\kappa_y}$. In each of the above welfare expressions, the first and third terms capture the expected welfare of patients for whom the physician does not use AI, whereas the second term represents the expected welfare of patients for whom the physician uses AI.

(a) Comparing the equilibrium accuracies gives $\rho^m \le \rho_x^*$, so it suffices to prove $\rho_y^* \le \rho^m$.

Define Δ as the difference between the AI firm's optimal expected profit under a disparate and an equal-accuracy algorithm: $\Delta := \pi^{\mathrm{D}}(\rho_x^*, \rho_y^*) - \pi^{\mathrm{E}}(\rho^m)$, with the profit functions in eqs. (1) and (2) evaluated at the equilibrium accuracies above. Let $\widehat{\ell}$ be the value satisfying $\rho_y^* = \rho^m$, and we can derive $\widehat{\ell} = \frac{b(\kappa_y - \beta\kappa_x)}{\beta(\kappa_x+\kappa_y)\theta}$.

Because ρ_x^* and ρ^m are independent of ℓ, and ρ_y^* is increasing in ℓ, we only need to prove $\widehat{\Delta} := \Delta|_{\ell=\widehat{\ell}} < 0$. Substituting $\ell = \widehat{\ell}$ and differentiating,

$$\widehat{\Delta} = \frac{f(\kappa_y - \beta\kappa_x)(f((2+\beta)\kappa_x + \kappa_y) - \kappa_x(\kappa_x + \kappa_y))}{\kappa_x(\kappa_x+\kappa_y)^2},$$

$$\frac{\partial \widehat{\Delta}}{\partial f} = -\frac{(\kappa_y - \beta\kappa_x)(\kappa_x(\kappa_x + \kappa_y) - 2f((2+\beta)\kappa_x + \kappa_y))}{\kappa_x(\kappa_x + \kappa_y)^2}.$$

Since $\frac{\partial \widehat{\Delta}}{\partial f}$ is increasing in f, $\widehat{\Delta}$ is convex in f. Because $\widehat{\Delta}|_{f=0} = 0$ and direct substitution yields $\widehat{\Delta}|_{f=\kappa_x/2} < 0$, convexity implies $\widehat{\Delta} < 0$ on the entire admissible range $0 < f < \frac{\kappa_x}{2}$. Finally, because Δ is decreasing in ℓ and $\widehat{\Delta} < 0$, the firm abandons the disparate design strictly before ℓ reaches $\widehat{\ell}$; since ρ_y^* is increasing in ℓ with $\rho_y^*|_{\ell=\widehat{\ell}} = \rho^m$, it follows that $\rho_y^* < \rho^m$ whenever the disparate design is optimal, completing the proof of (a).

(b) We first restate this part formally, writing $\alpha_x^{\pm} := \alpha^{\pm}(\rho_x^*)$ and $\alpha_m^{\pm} := \alpha^{\pm}(\rho^m)$: under a disparate algorithm the physician uses AI for type-x patients exactly when $\alpha \in (\alpha_x^-, \alpha_x^+)$ and for type-y patients exactly when $\alpha \in (\alpha_y^-, \alpha_y^+)$, and under the mandate she uses AI for both types exactly when $\alpha \in (\alpha_m^-, \alpha_m^+)$.

Comparing these cutoffs yields the ordering

$$\alpha_x^- < \alpha_m^- < \alpha_y^- < \alpha_y^+ < \alpha_m^+ < \alpha_x^+.$$

Moreover, each patient-indifference point sits a wedge $\frac{\theta r}{b}$ inside the corresponding use cutoff, the wedge being the physician's private benefit: $(1-\alpha)b = \rho_x^* b - c$ at $\alpha = \alpha_x^- + \frac{\theta r}{b}$, $\alpha b = \rho_x^* b - c$ at $\alpha = \alpha_x^+ - \frac{\theta r}{b}$, and likewise for ρ^m. The formal statement is:

The expected patient welfare is affected as follows. For type-x patients, welfare remains unchanged if $\alpha < \alpha_x^-$ *or* $\alpha > \alpha_x^+$*; improves if* $\alpha_x^- < \alpha < \min\{\alpha_m^-, \alpha_x^- + \frac{\theta r}{b}\}$ *or* $\max\{\alpha_m^+, \alpha_x^+ - \frac{\theta r}{b}\} < \alpha < \alpha_x^+$*; and worsens if* $\min\{\alpha_m^-, \alpha_x^- + \frac{\theta r}{b}\} < \alpha < \max\{\alpha_m^+, \alpha_x^+ - \frac{\theta r}{b}\}$*. For type-y patients, welfare remains unchanged if* $\alpha < \alpha_m^-$ *or* $\alpha > \alpha_m^+$*; worsens if* $\alpha_m^- < \alpha < \min\{\alpha_y^-, \alpha_m^- + \frac{\theta r}{b}\}$ *or* $\max\{\alpha_y^+, \alpha_m^+ - \frac{\theta r}{b}\} < \alpha < \alpha_m^+$*; and improves if* $\min\{\alpha_y^-, \alpha_m^- + \frac{\theta r}{b}\} < \alpha < \max\{\alpha_y^+, \alpha_m^+ - \frac{\theta r}{b}\}$.

We verify the statement interval by interval. For type-x patients, the mandate can only remove AI use: the use interval shrinks from (α_x^-, α_x^+) to (α_m^-, α_m^+). Outside (α_x^-, α_x^+) the physician never uses AI, so welfare is unchanged. On (α_x^-, α_m^-), where use stops, welfare changes from $\rho_x^* b - c$ to $(1-\alpha)b$, an improvement precisely when $\alpha < \alpha_x^- + \frac{\theta r}{b}$; on (α_m^+, α_x^+) it changes from $\rho_x^* b - c$ to αb, an improvement precisely when $\alpha > \alpha_x^+ - \frac{\theta r}{b}$. On the middle interval, where use continues, welfare falls from $\rho_x^* b - c$ to $\rho^m b - c \le \rho_x^* b - c$. Intersecting each comparison with the interval on which it applies yields exactly the type-x intervals in the statement, with each min and max recording whether the indifference point falls inside

that interval.

For type-y patients the pattern reverses: the mandate can only add AI use, the use interval expanding from (α_y^-, α_y^+) to (α_m^-, α_m^+). Outside (α_m^-, α_m^+) welfare is unchanged. On (α_m^-, α_y^-), where use begins, welfare changes from $(1-\alpha)b$ to $\rho^m b - c$, an improvement precisely when $\alpha > \alpha_m^- + \frac{\theta r}{b}$; on (α_y^+, α_m^+) it changes from αb to $\rho^m b - c$, an improvement precisely when $\alpha < \alpha_m^+ - \frac{\theta r}{b}$. Where use already occurred, welfare rises from $\rho_y^* b - c$ to $\rho^m b - c \geq \rho_y^* b - c$ by part (a). Intersecting as before yields the type-y intervals in the statement.

(c) We prove this part using Proposition 1. Substituting the equilibrium accuracies stated below eq. (A6) into the welfare expressions there, we obtain $W_x(\rho^m) > W_x(\rho_x^*) \iff f < f_x$, where $f_x := \frac{2c\kappa_x(\kappa_x+\kappa_y)}{b((2+\beta)\kappa_x+\kappa_y)}$. Analogously, when $\ell \to 0$, we have $W_y(\rho^m) > W_y(\rho_y^*)|_{\ell\to 0} \iff f > f_y$, where $f_y := \frac{2c\kappa_y(\kappa_x+\kappa_y)}{b((2\beta+1)\kappa_y+\beta\kappa_x)}$. We can verify that $f_x < f_y$. It is straightforward to verify $\Delta|_{\ell\to 0} > 0$: the disparate design strictly dominates near $\ell = 0$, so we may drop the parameter restrictions of Proposition 1 that otherwise guarantee profit dominance ($\beta(\beta+2)\kappa_x > \kappa_y$ and the associated lower bound on f). Write $f_\pi(\ell) := \frac{b\kappa_y(2c-2\theta r+\theta\ell)}{\beta(b+\theta\ell)^2}$ for the type-y participation bound from Proposition 1. Then,

$$
\begin{aligned}
f_x > f_\pi(0) &\iff c < \frac{\kappa_y((2+\beta)\kappa_x+\kappa_y)\theta r}{\kappa_y(2\kappa_x+\kappa_y)-\beta\kappa_x^2} \\
f_y < \frac{\kappa_x}{2} &\iff b > \frac{4c\kappa_y(\kappa_x+\kappa_y)}{\kappa_x(\kappa_y+\beta(\kappa_x+2\kappa_y))} \\
f_y < \frac{b\kappa_y}{2\beta(b+\theta\ell)} &\iff b > \frac{4\beta c(\kappa_x+\kappa_y)}{\kappa_y+\beta(\kappa_x+2\kappa_y)}.
\end{aligned}
$$

We can verify that $\frac{4c\kappa_y(\kappa_x+\kappa_y)}{\kappa_x(\kappa_y+\beta(\kappa_x+2\kappa_y))} > \frac{4\beta c(\kappa_x+\kappa_y)}{\kappa_y+\beta(\kappa_x+2\kappa_y)}$. Therefore, suppose the two preconditions stated in part (c) hold; then all three inequalities above are satisfied. In the limit $\ell \to 0$, mandating equal accuracy yields three regimes:

- if $f_\pi(0) < f < f_x$, then $W_x(\rho^m) > W_x(\rho_x^*)$ while $W_y(\rho^m) < W_y(\rho_y^*)$;
- if $f_x < f < f_y$, then both types are worse off under the mandate, that is, $W_x(\rho^m) < W_x(\rho_x^*)$ and $W_y(\rho^m) < W_y(\rho_y^*)$ (each endpoint leaves one type exactly indifferent);
- if $f_y < f < \min\{\frac{\kappa_x}{2}, \frac{b\kappa_y}{2\beta(b+\theta\ell)}\}|_{\ell\to 0}$, then $W_x(\rho^m) < W_x(\rho_x^*)$ while $W_y(\rho^m) > W_y(\rho_y^*)$.

Every welfare expression and equilibrium condition above is continuous in ℓ. Hence, for each parameter tuple satisfying these conditions there exists $\varepsilon > 0$ such that, for all $0 \leq \ell < \varepsilon$,

the same three regimes persist, with the boundaries evaluated at that ℓ and the type-y indifference point replaced by a threshold $\bar{f}$, defined by $W_y(\rho^m) = W_y(\rho_y^*)$ (so that $\bar{f} \to f_y$ as $\ell \to 0$), exactly as stated in part (c) of the proposition. This completes the proof. *Q.E.D.*

References

Abràmoff MD, Cunningham B, Patel B, Eydelman MB, Leng T, Sakamoto T, Blodi B, Grenon SM, Wolf RM, Manrai AK, et al. (2022a) Foundational considerations for artificial intelligence using ophthalmic images. *Ophthalmology* 129(2):e14–e32.

Abràmoff MD, Dai T, Zou J (2024) Scaling adoption of medical AI-reimbursement from value-based care and fee-for-service perspectives. *NEJM AI* 1(5):AIpc2400083.

Abràmoff MD, Roehrenbeck C, Trujillo S, Goldstein J, Graves AS, Repka MX, Silva III E (2022b) A reimbursement framework for artificial intelligence in healthcare. *npj Digital Medicine* 5(1).

Balakrishnan M, Ferreira KJ, Tong J (2026) Human-algorithm collaboration with private information: Naïve advice-weighting behavior and mitigation. *Management Science* 72(1):265–284.

Benjamin R (2016) Innovating inequity: If race is a technology, postracialism is the genius bar. *Ethnic and Racial Studies* 39(13):2227–2234.

Cheatham M (2024) Who Pays for Healthcare AI—Usage-based Models (Part II). https://morgancheatham.substack.com/p/who-pays-for-healthcare-ai-part-ii, accessed: February 13, 2024.

Chen IY, Johansson FD, Sontag D (2018) Why is my classifier discriminatory? *Advances in Neural Information Processing Systems 31*, 3543–3554.

Chen Y, Li J, Zhang J (2022) Efficient liability in expert markets. *International Economic Review* 63(4):pp. 1717–1744.

CMS (2022) Nondiscrimination in health programs and activities. https://bit.ly/fed-gov-non-discrim-hlth, comments close on October 3, 2022. Docket ID: HHS-OS-2022-0012. Document Number: 2022-16217. Document Citation: 87 FR 47824. Pages: 47824-47920.

CMS (2024) Nondiscrimination in health programs and activities. https://bit.ly/fed-gov-non-discrim-hlth-24, effective date: July 5, 2024. RIN: 0945-AA17. Document Number: 2024-08711. Document Citation: 89 FR 37522. Pages: 37522-37703.

Corbett-Davies S, Goel S (2018) The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023* .

Corbett-Davies S, Pierson E, Feller A, Goel S, Huq A (2017) Algorithmic decision making and the cost of fairness. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*, 797–806 (New York, NY, USA: ACM), URL http://dx.doi.org/10.1145/3097983.3098095.

Currie J, MacLeod WB (2008) First do no harm? Tort reform and birth outcomes. *Quarterly Journal of Economics* 123(2):795–830.

Dai T, Singh S (2020) Conspicuous by its absence: Diagnostic expert testing under uncertainty. *Marketing Science* 39(3):540–563.

Dai T, Singh S (2025) Artificial intelligence on call: The physician's decision of whether to use AI in clinical practice. *Journal of Marketing Research* 62(5):854–875.

Darby MR, Karni E (1973) Free competition and the optimal amount of fraud. *Journal of Law and Economics* 16(1):67–88.

Diao W, Harutyunyan M, Jiang B (2023) Consumer fairness concerns and dynamic pricing in a channel. *Marketing Science* 42(3):569–588.

Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Management Science* 64(3):1155–1170.

Dulleck U, Kerschbamer R, Sutter M (2011) The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review* 101(2):526–55.

FDA (2025) Artificial intelligence and machine learning (AI/ML)-enabled medical devices. https://bit.ly/FDA-AI-ML, published by the U.S. Food and Drug Administration. Accessed: January 19, 2026.

Fong Y, Liu T (2018) Liability and reputation in credence goods markets. *Economics Letters* 166:35–39.

Fu R, Aseri M, Singh PV, Srinivasan K (2022) "Un" fair machine learning algorithms. *Management Science* 68(6):4173–4195.

Gianfrancesco MA, Tamang S, Yazdany J, Schmajuk G (2018) Potential biases in machine learning algorithms using electronic health record data. *JAMA Internal Medicine* 178(11):1544–1547.

Gichoya JW, Thomas K, Celi LA, Safdar N, Banerjee I, Banja JD, Seyyed-Kalantari L, Trivedi H, Purkayastha S (2023) AI pitfalls and what not to do: Mitigating bias in AI. *British Journal of Radiology* 96(1150).

Gillis T, McLaughlin B, Spiess J (2021) On the fairness of machine-assisted human decisions. *arXiv preprint arXiv:2110.15310* .

Goodman KE, Morgan DJ, Hoffmann DE (2023) Clinical algorithms, antidiscrimination laws, and medical device regulation. *JAMA* 329(4):285–286.

Gottlieb S (2024) Congress must update FDA regulations for medical AI. *JAMA Health Forum* 5(7):e242691.

Green G, Defoe EC (1978) What is a clinical algorithm? *Clinical Pediatrics* 17(5):457–463.

Guan X, Cao H, Li KJ, Ding Y (2024) Product safety and liability with deceptive advertising and moral hazard. *Marketing Science* 44(2):287–305.

Heckler M (1985) Report of the secretary's task force report on Black and minority health volume I: Executive summary. Technical report, Government Printing Office, Washington, DC.

Israeli A (2018) Online map enforcement: evidence from a quasi-experiment. *Marketing Science* 37(5):710–732.

Iyer G, Ke TT (2024) Competitive model selection in algorithmic targeting. *Marketing Science* 43(6):1226–1241.

Iyer G, Singh S (2018) Voluntary product safety certification. *Management Science* 64(2):695–714.

Ke TT, Sudhir K (2023) Privacy rights and data security: GDPR and personal data markets. *Management Science* 69(8):4389–4412.

Lambrecht A, Tucker C (2024) Apparent algorithmic discrimination and real-time algorithmic learning in digital search advertising. *Quantitative Marketing and Economics* 1–31.

Leong A, Wang J, Wolf R, Channa R, Abramoff MD, Lehmann H, Liu TA (2023) Autonomous artificial intelligence (AI) increases health equity for patients who are more at risk for poor visual outcomes due to diabetic eye disease (DED). *Investigative Ophthalmology & Visual Science* 64(8):243–243.

Leung E, Paolacci G, Puntoni S (2018) Man versus machine: Resisting automation in identity-based consumer behavior. *Journal of Marketing Research* 55(6):818–831.

Li X, Li KJ (2023) Beating the algorithm: Consumer manipulation, personalized pricing, and big data management. *Manufacturing & Service Operations Management* 25(1):36–49.

Liaw W, Kueper JK, Lin S, Bazemore A, Kakadiaris I (2022) Competencies for the use of artificial intelligence in primary care. *The Annals of Family Medicine* 20(6):559–563.

Liu LT, Dean S, Rolf E, Simchowitz M, Hardt M (2018) Delayed impact of fair machine learning. *International Conference on Machine Learning*, 3150–3158 (PMLR).

Longoni C, Bonezzi A, Morewedge CK (2019) Resistance to medical artificial intelligence. *Journal of Consumer Research* 46(4):629–650.

Luo X, Tong S, Fang Z, Qu Z (2019) Frontiers: Machines vs. humans: The impact of artificial intelligence chatbot disclosure on customer purchases. *Marketing Science* 38(6):937–947.

Margolis CZ (1983) Uses of clinical algorithms. *JAMA* 249(5):627.

McLaughlin B, Spiess J (2022) Algorithmic assistance with recommendation-dependent preferences. *arXiv preprint arXiv:2208.07626* .

Mehrabi N, Morstatter F, Saxena N, Lerman K, Galstyan A (2021) A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54(6):1–35.

Mello MM, Roberts JL (2024) Antidiscrimination law meets artificial intelligence: New requirements for health care organizations and insurers. *JAMA Health Forum* 5(8):e243397–e243397.

Mohammadi B, Malik N, Derdenger T, Srinivasan K (2024) Regulating explainable artificial intelligence (XAI) may harm consumers. *Marketing Science* 44(3):711–724.

Nelson AR (2002) Unequal treatment: Confronting racial and ethnic disparities in health care. *Journal of the National Medical Association* 94(8):666–668.

Obermeyer Z, Nissan R, Stern M, Eaneff S, Bembeneck EJ, Mullainathan S (2021) Algorithmic bias playbook. *Center for Applied AI at Chicago Booth* 7–8.

Obermeyer Z, Powers B, Vogeli C, Mullainathan S (2019) Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366(6464):447–453.

Parikh RB, Helmchen LA (2022) Paying for artificial intelligence in medicine. *NPJ digital medicine* 5(1):63.

Price WN, Gerke S, Cohen IG (2019) Potential liability for physicians using artificial intelligence. *JAMA* 322(18):1765–1766.

Rajpurkar P, Chen E, Banerjee O, Topol EJ (2022) AI in health and medicine. *Nature Medicine* 28(1):31–38.

Samorani M, Harris SL, Blount LG, Lu H, Santoro MA (2022) Overbooked and overlooked: Machine learning and racial bias in medical appointment scheduling. *Manufacturing & Service Operations Management* 24(6):2825–2842.

Schubert T, Oosterlinck T, Stevens RD, Maxwell PH, van der Schaar M (2025) AI education for clinicians. *EClinicalMedicine* 79.

Schuitmaker L, Drogt J, Benders M, Jongsma K (2025) Physicians' required competencies in AI-assisted clinical settings: A systematic review. *British Medical Bulletin* 153(1):ldae025.

Shimao H, Khern-am nuai W, Kannan K, Cohen MC (2022) Strategic best response fairness in fair machine learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 664–664.

Tipton K, Leas BF, Flores E, Jepson C, Aysola J, Cohen J, Harhay M, Schmidt H, Weissman G, Treadwell J, et al. (2023) Impact of healthcare algorithms on racial and ethnic disparities in health and healthcare. *Agency for Healthcare Research and Quality (US)* .

Topol EJ (2019) High-performance medicine: The convergence of human and artificial intelligence. *Nature Medicine* 25(1):44–56.

White House (2023) Executive order on the safe, secure, and trustworthy development and use of artificial intelligence. https://docs.house.gov/meetings/FD/FD00/20240206/116793/HHRG-118-FD00-20240206-SD001.pdf, accessed: March 28, 2026.

Wu K, Wu E, Theodorou B, Liang W, Mack C, Glass L, Sun J, Zou J (2023) Characterizing the clinical adoption of medical AI devices through U.S. insurance claims. *NEJM AI* 1(1), ISSN 2836-9386.

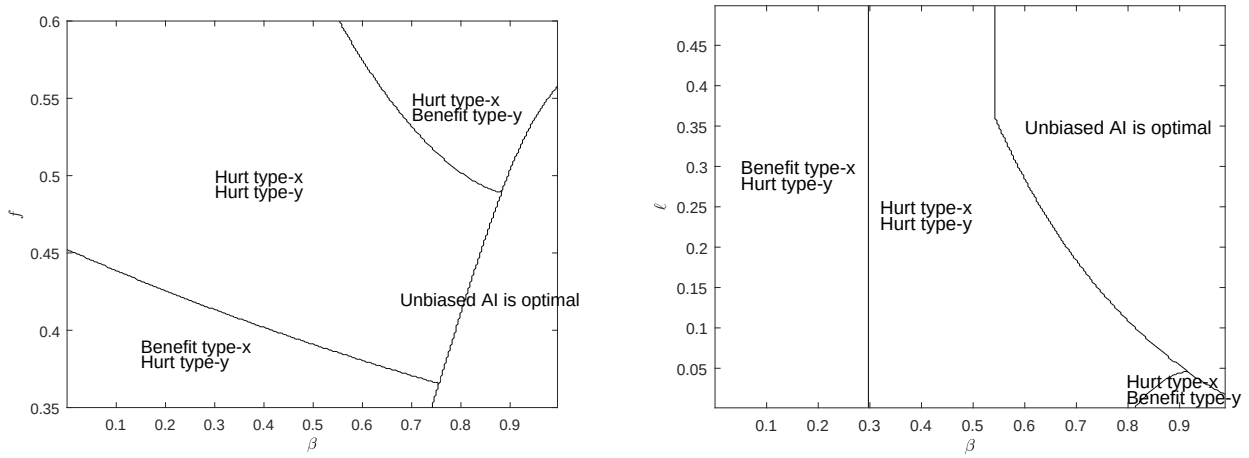
Zimmermann L, Somasundaram J, Saha B (2024) Adoption of new technology vaccines. *Journal of Marketing* 88(4):1–21.

Online Appendix for "Algorithm Design and Physician Liability"

This Online Appendix is organized as follows. The first part extends the welfare comparisons under equal-accuracy requirements. The second part provides technical derivations for the endogenous-pricing extension. The third part reports the welfare-maximizing-liability computations. The fourth part develops the extension with type-specific priors and its additional comparative statics.

OA1. Effect of Mandating Equal Accuracy on Patient Welfare (Section 5.3)

To assess how an equal-accuracy requirement affects aggregate welfare over a broader parameter range, we report numerical comparisons across (β, ℓ, f) in Figure OA1.



(a) Variation in (β, f)

(b) Variation in (ℓ, β)

Figure OA1: Welfare effects of an equal-accuracy requirement for each patient type. Regions labeled "Unbiased AI is optimal" are those in which the equal-accuracy design is optimal.

Figure OA1a complements Proposition 4(c). As f increases, the welfare effect of the equal-accuracy mandate typically transitions from harming only type-y patients, to harming both types, and then to harming only type-x patients. Intuitively, higher f strengthens the AI firm's incentive to invest in accuracy for type-y patients under the equal-accuracy requirement; once these gains become sufficiently large, the probability that type-y patients are harmed falls.

The same figure shows how these welfare patterns vary with the relative size of the disadvantaged segment. As β increases, the equal-accuracy requirement again tends to move from harming only type-y patients, to harming both types, and eventually to harming primarily type-x patients. When β is large, the AI firm internalizes a larger market on the

type-y side, which increases the return to improving their accuracy under the requirement and limits the scope for welfare loss in that group. When β is small, the requirement can instead improve welfare for type-x patients by tempering overuse on the margin, partially offsetting the loss from reduced accuracy.

Finally, Figure OA1b shows that higher liability can increase the likelihood that type-y patients are harmed by the equal-accuracy requirement. When ℓ is large, the disparate-design equilibrium already induces substantial investment in type-y accuracy, reducing the incremental benefit of further improvements under the requirement, while utilization effects remain. As a result, combining a high-liability environment with an equal-accuracy requirement—two instruments intended to protect disadvantaged patients—can be counterproductive in equilibrium.

OA2. AI Firm's Pricing Decision (Section 6.1)

Proof of Proposition 5. The AI firm's profit functions under disparate and equal-accuracy designs are given in eqs. (1) and (2). We first consider the case in which the firm offers a disparate design. The Hessian matrix $\mathbb{H}^{\mathrm{D}}$ of the profit function $\pi^{\mathrm{D}}(\rho_x, \rho_y, f)$ is given by:

$$\mathbb{H}^{\mathrm{D}} := \begin{bmatrix} -\frac{4(1+\beta)}{b} & 2 & \frac{2\beta(b+\theta\ell)}{b} \\ 2 & -2\kappa_x & 0 \\ \frac{2\beta(b+\theta\ell)}{b} & 0 & -2\kappa_y \end{bmatrix} \tag{OA1}$$

The corresponding leading principal minors of $\mathbb{H}^{\mathrm{D}}$ are evaluated as:

$$\mathbb{D}_1 = -\frac{4(1+\beta)}{b}, \quad \mathbb{D}_2 = \frac{8(1+\beta)\kappa_x}{b} - 4 \text{ and } \mathbb{D}_3 = \frac{4D(\ell)}{b^2},$$

where

$$D(\ell) := 2b^2(\beta^2\kappa_x + \kappa_y) - 4b\kappa_x\Big((\beta+1)\kappa_y - \beta^2\theta\ell\Big) + 2\beta^2\kappa_x\theta^2\ell^2.$$

To ensure the profit function $\pi^{\mathrm{D}}(\rho_x, \rho_y, f)$ is jointly concave in (f, ρ_x, ρ_y), the Hessian matrix must be negative definite, which requires $\mathbb{D}_1 < 0$, $\mathbb{D}_2 > 0$, and $\mathbb{D}_3 < 0$. These conditions yield the parameter space boundaries $b < \frac{2(1+\beta)\kappa_x\kappa_y}{\beta^2\kappa_x+\kappa_y}$ and $\ell < \sqrt{\frac{2b\kappa_x\kappa_y - b^2\kappa_y + 2b\beta\kappa_x\kappa_y}{\beta^2\theta^2\kappa_x}} - \frac{b}{\theta}$.

Under these concavity conditions, we derive the optimal decision vector (f, ρ_x, ρ_y). The

first-order condition $\frac{\partial \pi^{\mathrm{D}}(\rho_x,\rho_y,f)}{\partial f} = 0$ yields the unique best-response price

$$f(\rho_x,\rho_y) = \frac{b\Big(\beta(2\rho_y-1)+2\rho_x-1\Big) - 2\theta\beta\ell(1-\rho_y) + 2(\beta+1)\theta r}{4(\beta+1)}.$$

Substituting $f(\rho_x,\rho_y)$ back into π^{D} and imposing the first-order conditions with respect to ρ_x and ρ_y yields the optimal accuracies (ρ_x^*,ρ_y^*) and the induced optimal price $f^{\mathrm{D}} := f(\rho_x^*,\rho_y^*)$:

$$\rho_x^* = \frac{1}{2} + \frac{b\kappa_y\theta\Big(\beta\ell - 2(\beta+1)r\Big)}{D(\ell)}, \qquad \rho_y^* = \frac{1}{2} + \frac{\beta\kappa_x\theta(b+\theta\ell)\Big(\beta\ell - 2(\beta+1)r\Big)}{D(\ell)},$$

and

$$f^{\mathrm{D}} = \frac{b\kappa_x\kappa_y\theta\Big(\beta\ell - 2(\beta+1)r\Big)}{D(\ell)}.$$

To simplify exposition, we define the auxiliary variable $\mathcal{T} = \frac{\theta(\beta\ell-2(\beta+1)r)}{D(\ell)}$, allowing the equilibrium expressions to be rewritten compactly as $\rho_x^* = \frac{1}{2} + b\kappa_y\mathcal{T}$, $\rho_y^* = \frac{1}{2} + \beta\kappa_x(b+\theta\ell)\mathcal{T}$, and $f^{\mathrm{D}} = b\kappa_x\kappa_y\mathcal{T}$.

We next identify the parameter regimes that guarantee the validity of these interior optimal solutions. First, given $D(\ell) < 0$, the requirements $\rho_x^* > 1/2$, $\rho_y^* > 1/2$, and $f^{\mathrm{D}} > 0$ hold if and only if $\mathcal{T} > 0$, that is, if and only if $\ell < \frac{2(1+\beta)r}{\beta}$. Additionally, imposing the logical upper bounds on accuracy levels ($\rho_x^* < 1$ and $\rho_y^* < 1$) requires $\mathcal{T} < \frac{1}{2b\kappa_y}$ and $\mathcal{T} < \frac{1}{2\beta\kappa_x(b+\theta\ell)}$, respectively. To guarantee that $f^{\mathrm{D}} > \theta r$, we require $\mathcal{T} > \frac{\theta r}{b\kappa_x\kappa_y}$. Finally, substituting $(\rho_y^*, f^{\mathrm{D}})$ into the profit function isolated to type-y patients yields $\pi_y(\rho_y^*, f^{\mathrm{D}}) = y_1\mathcal{T}^2 + y_2\mathcal{T}$, where:

$$y_1 = \beta\kappa_x^2\kappa_y\left(b^2\beta - 2b\kappa_y + 2b\beta\theta\ell + \beta\theta^2\ell^2\right) \quad \text{and} \quad y_2 = -\beta\kappa_x\kappa_y(\ell-2r)\theta.$$

To ensure a well-behaved optimization problem, we impose $y_1 < 0$, which holds if and only if $\kappa_y > \frac{b^2\beta+2b\beta\theta\ell+\beta\theta^2\ell^2}{2b}$ (the standing assumption $\kappa_y > \kappa_x$ is maintained throughout). Setting $y_2 > 0$ yields the condition $\ell < 2r$. Under these restrictions, the unique non-zero root of $\pi_y(\rho_y^*, f^{\mathrm{D}}) = 0$ with respect to $\mathcal{T}$ is strictly positive:

$$\mathcal{T} = \frac{(\ell-2r)\theta}{\kappa_x\left(b^2\beta - 2b\kappa_y + 2b\beta\theta\ell + \beta\theta^2\ell^2\right)} > 0.$$

Consequently, ensuring positive profits from the type-y segment ($\pi_y(\rho_y^*, f^{\mathrm{D}}) > 0$) is mathemat-

ically equivalent to bounding the auxiliary variable such that $\mathcal{T} < \frac{(\ell-2r)\theta}{\kappa_x(b^2\beta-2b\kappa_y+2b\beta\theta\ell+\beta\theta^2\ell^2)}$.

Next consider the equal-accuracy design, imposing $\rho_x = \rho_y \equiv \rho$. The Hessian matrix $\mathbb{H}^{\mathrm{E}}$ of the profit function $\pi^{\mathrm{E}}(\rho, f)$ is given by:

$$\mathbb{H}^{\mathrm{E}} := \begin{bmatrix} -\frac{4(1+\beta)}{b} & 2(1+\beta) \\ 2(1+\beta) & -2(\kappa_x+\kappa_y) \end{bmatrix} \tag{OA2}$$

The corresponding leading principal minors of $\mathbb{H}^{\mathrm{E}}$ are evaluated as:

$$\mathbb{D}_1 = -\frac{4(1+\beta)}{b} \text{ and } \mathbb{D}_2 = -\frac{4(1+\beta)(b(1+\beta)-2(\kappa_x+\kappa_y))}{b}.$$

To ensure the profit function $\pi^{\mathrm{E}}(\rho, f)$ is jointly concave in (f, ρ), the Hessian matrix must be negative semi-definite, which requires $\mathbb{D}_1 \leq 0$ and $\mathbb{D}_2 \geq 0$, which yield $b < \frac{2(\kappa_x+\kappa_y)}{1+\beta}$. Under these concavity conditions, we derive the optimal decision vector (f, ρ). The first-order condition $\frac{\partial \pi^{\mathrm{E}}(\rho,f)}{\partial f} = 0$ gives

$$f(\rho) = \left(\frac{\rho}{2} - \frac{1}{4}\right)b + \frac{1}{2}\theta r.$$

Substituting into π^{E} and imposing $\frac{\partial \pi^{\mathrm{E}}(\rho,f(\rho))}{\partial \rho} = 0$ yields

$$\rho^* = \frac{1}{2} + \frac{(\beta+1)\theta r}{2(\kappa_x+\kappa_y)-b(\beta+1)}, \qquad f^{\mathrm{E}} := f(\rho^*) = \frac{(\kappa_x+\kappa_y)\theta r}{2(\kappa_x+\kappa_y)-b(\beta+1)}.$$

Ensuring that the equilibrium accuracy levels are interior, $1/2 < \rho^* < 1$, and that the condition $f^{\mathrm{E}} > \theta r$ holds yields the following joint parameter restrictions:

$$r < \frac{\kappa_x+\kappa_y}{2\theta(1+\beta)} \quad \text{and} \quad \frac{\kappa_x+\kappa_y}{1+\beta} < b < \frac{2(\kappa_x+\kappa_y)-2r\theta(1+\beta)}{1+\beta}.$$

In addition, requiring $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*, f^{\mathrm{D}}) > \pi^{\mathrm{E}}(\rho^*, f^{\mathrm{E}})$ leads to (the left-hand side below equals $4\left[\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*, f^{\mathrm{D}}) - \pi^{\mathrm{E}}(\rho^*, f^{\mathrm{E}})\right]$):

$$\begin{aligned} &4b\kappa_x\kappa_y\left(b\left(\beta^2\kappa_x+\kappa_y\right)-2(1+\beta)\kappa_x\kappa_y\right)\mathcal{T}^2 \\ &+4\theta\kappa_x\kappa_y\mathcal{T}\left(2r(1+\beta)+\beta\ell(2b\beta\kappa_x\mathcal{T}-1)\right) \\ &+4\theta^2\left(\frac{(1+\beta)(\kappa_x+\kappa_y)r^2}{b\left(b(1+\beta)-2(\kappa_x+\kappa_y)\right)}+\beta^2\kappa_x^2\kappa_y\ell^2\mathcal{T}^2\right) > 0. \end{aligned}$$

Taken together, the conditions above characterize when a disparate AI design is optimal; see Figure OA2b for a numerical illustration.

We now compare the maximized payoffs as ℓ varies. By the envelope theorem, the maximized payoff under the equal-accuracy design, $\pi^{\mathrm{E}}(\rho^*, f^{\mathrm{E}})$, is independent of ℓ. Under the disparate design, the envelope theorem implies

$$\frac{\partial \pi^{\mathrm{D}}(\rho_x^*, \rho_y^*, f^{\mathrm{D}})}{\partial \ell} = \left.\frac{\partial \pi^{\mathrm{D}}(\rho_x, \rho_y, f)}{\partial \ell}\right|_{(\rho_x,\rho_y,f)=(\rho_x^*,\rho_y^*,f^{\mathrm{D}})} = -\frac{2\beta\theta f^{\mathrm{D}}}{b}\,(1-\rho_y^*) < 0,$$

where the inequality holds within the feasible region we discussed earlier; thus, the maximized disparate-design payoff is strictly decreasing in ℓ. Moreover, as $\ell \to 0$, the disparate-design problem converges to the environment without the liability channel, and allowing (ρ_x, ρ_y) to differ is (weakly) valuable under the cost asymmetry. In particular, the disparate-design optimum yields a strictly higher payoff than the best equal-accuracy design for ℓ near zero. Because $\pi^{\mathrm{E}}(\rho^*, f^{\mathrm{E}})$ is constant whereas $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*, f^{\mathrm{D}})$ declines in ℓ, there exists at most one cutoff $\ell^\dagger$ within the feasible region such that the firm prefers a disparate design for $\ell < \ell^\dagger$. If no such $\ell^\dagger$ falls within the feasible region, the disparate design remains globally optimal in the region. Conversely, when $\ell \geq \ell^\dagger$, the equal-accuracy design can become optimal; see Figure OA2b for a numerical illustration.

We next show that whenever a disparate design is optimal, it must satisfy $\rho_x^* > \rho_y^*$. At $\ell \to 0$, the expressions above imply $\rho_x^* > \rho_y^*$. Indeed, $\rho_x^* - \rho_y^* = \mathcal{T}\Big(b\kappa_y - \beta\kappa_x(b + \theta\ell)\Big)$; with $\mathcal{T} > 0$, the bracket is strictly decreasing in ℓ and positive at $\ell = 0$ because $\kappa_y > \beta\kappa_x$, so the gap changes sign at most once. Suppose instead that for some $\hat{\ell}$ the disparate-design optimum satisfies $\rho_x^* = \rho_y^*$. Then the candidate solution is itself equal-accuracy, and the firm can (weakly) improve by switching to the equal-accuracy problem; moreover, because $\pi^{\mathrm{D}}(\rho_x^*, \rho_y^*, f^{\mathrm{D}})$ is decreasing in ℓ while $\pi^{\mathrm{E}}(\rho^*, f^{\mathrm{E}})$ is independent of ℓ, once $\rho_x^* = \rho_y^*$ holds at some $\hat{\ell}$, the equal-accuracy design is optimal at $\hat{\ell}$ and for all larger ℓ. Hence, if the disparate design is optimal, it must be that $\rho_x^* > \rho_y^*$.

Finally, ρ_x^* (and similarly f^{D}) can be non-monotone in ℓ. Differentiating yields

$$\frac{\partial \rho_x^*}{\partial \ell} = \frac{b\beta\kappa_y\theta \cdot E_1(\ell)}{2\Big(b^2(\beta^2\kappa_x + \kappa_y) - 2b\kappa_x\Big((\beta+1)\kappa_y - \beta^2\theta\ell\Big) + \beta^2\kappa_x\theta^2\ell^2\Big)^2},$$

where

$$E_1(\ell) = b^2(\beta^2\kappa_x + \kappa_y) - 2b(\beta+1)\kappa_x(\kappa_y - 2\beta\theta r) + \beta\kappa_x\theta^2\ell\Big(4(\beta+1)r - \beta\ell\Big). \tag{OA3}$$

Because $E_1(\ell)$ is concave in ℓ, there exist parameter values for which $E_1(0) < 0$ but $E_1(\epsilon) > 0$ for some small $\epsilon > 0$ below the minimum of the design-switch cutoff $\ell^\dagger$ and $\frac{4(\beta+1)r}{\beta}$. To establish this, it suffices to ensure that $b^2(\beta^2\kappa_x + \kappa_y) - 2b(\beta+1)\kappa_x(\kappa_y - 2\beta\theta r)$ is strictly negative but sufficiently close to zero, so that $E_1(\epsilon) > 0$ for a small $\epsilon > 0$. This implies that ρ_x^* initially decreases and subsequently increases with ℓ. The same logic applies to f^{D}. *Q.E.D.*

Proposition OA1. *There exist parameter values under which, as ℓ increases, equilibrium AI use for type-y patients first decreases and then increases.*

Proof of Proposition OA1. With endogenous pricing and a disparate design,

$$d_y^{\mathrm{D}} = \beta \cdot \frac{(2\rho_y^* - 1)b - 2(1-\rho_y^*)\theta\ell - 2f^{\mathrm{D}} + 2\theta r}{b},$$

where $(\rho_y^*, f^{\mathrm{D}})$ are given in Proposition 5. Differentiating d_y^{D} with respect to ℓ and simplifying yields

$$\frac{\partial d_y^{\mathrm{D}}}{\partial \ell} = \frac{\beta\kappa_y\theta\Big((\beta+2)\kappa_x - b\Big)\cdot E_1(\ell)}{\Big(b^2(\beta^2\kappa_x + \kappa_y) - 2b\kappa_x\Big((\beta+1)\kappa_y - \beta^2\theta\ell\Big) + \beta^2\kappa_x\theta^2\ell^2\Big)^2},$$

where $E_1(\ell)$ is defined in eq. (OA3) in the proof of Proposition 5 and it is a concave quadratic in ℓ. Let $\ell_{\min}$ denote the smaller root of $E_1(\ell) = 0$:

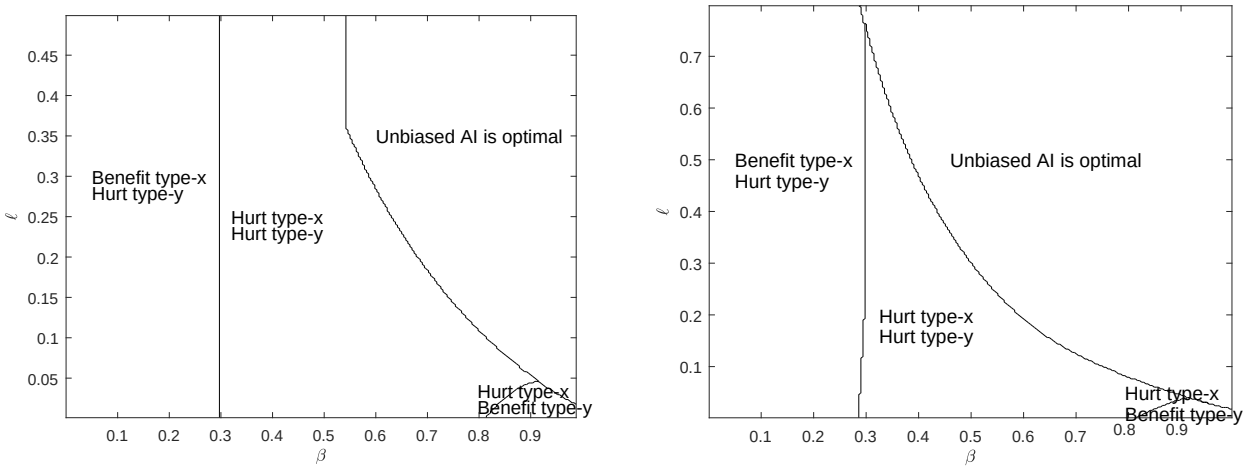
$$\ell_{\min} = \frac{2(\beta+1)r}{\beta} - \frac{\sqrt{b^2(\beta^2 + \kappa_y/\kappa_x) - 2b(\beta+1)\left(\kappa_y - 2\beta r\theta\right) + 4(\beta+1)^2r^2\theta^2}}{\beta\theta}.$$

If $(\beta+2)\kappa_x > b$, $E_1(0) < 0$, and $E_1(0) > -4(\beta+1)^2\kappa_x\theta^2r^2$ (so that E_1 admits a real root), then $\frac{\partial d_y^{\mathrm{D}}}{\partial \ell} < 0$ for ℓ near zero and becomes positive for ℓ just above $\ell_{\min}$. The last condition cannot be dispensed with: E_1 is concave with vertex at $\ell = 2(\beta+1)r/\beta$, which coincides with the boundary of the feasible region (where $\mathcal{T} > 0$ fails), so E_1 is strictly increasing throughout the feasible region and its maximum there equals $E_1(0) + 4(\beta+1)^2\kappa_x\theta^2r^2$. Consequently the crossing at $\ell_{\min}$ is unique and the sign change is global rather than local. Provided $\ell_{\min}$ lies below the design-switch cutoff $\ell^\dagger$ (so the firm still supplies a disparate design in this

neighborhood), the comparative static in Proposition OA1 follows. *Q.E.D.*

The mechanism behind Proposition OA1 parallels that of Proposition 2: liability changes the physician's marginal willingness to rely on AI for type-y patients, and the firm responds by jointly adjusting price and accuracy, which can reverse the direction of utilization as ℓ increases.

We next examine whether an equal-accuracy requirement can still reduce welfare for both patient types when the AI firm endogenizes the per-use price (that is, when f becomes a choice variable and we impose $c = f$). Figure OA2 reports the comparison. Panel (a) corresponds to the baseline model with exogenous (c, f), whereas panel (b) corresponds to the extension in which the firm chooses f and c moves one-for-one with it. The two panels use the same parameter values and differ only in whether (c, f) are fixed or endogenously determined. The qualitative welfare regions are unchanged: the central patterns from the baseline carry over, and the mandate can still generate the same welfare reversals.



(a) Exogenous f (b) Endogenous f

Figure OA2: Welfare effects of an equal-accuracy requirement under exogenous and endogenous pricing. Panel (a) reproduces the baseline with exogenous (c, f); panel (b) endogenizes f and imposes $c = f$. Regions labeled "Unbiased AI is optimal" are those in which the equal-accuracy design is optimal.

OA3. Welfare-Maximizing Liability (Section 6.2)

For each parameter tuple, we compute equilibrium aggregate welfare under the disparate-design and equal-accuracy-design regimes as functions of liability and then compare these values at the relevant design boundary. Operationally, we first identify the design-switch cutoff implied by Proposition 1, and then evaluate welfare on each side of that cutoff using the corresponding equilibrium design.

Across the numerical configurations reported in the main text, the welfare-maximizing liability is attained at the regime boundary: either at the largest ℓ for which the disparate design remains optimal or at the smallest ℓ that induces the equal-accuracy design. This procedure yields the two benchmark patterns illustrated in Figures 5a and 5b.

OA4. Type-Specific Priors (Section 6.3)

As in the baseline, the physician's AI-use rule is characterized by two prior cutoffs, and conditional on use the physician follows the AI signal.

Lemma OA1. *(a) For type-x patients, the physician uses AI if and only if*

$$\frac{(1-\rho_x)b + c - \theta r}{b} < \alpha_x < \min\left\{\overline{\alpha}, \frac{\rho_x b - c + \theta r}{b}\right\}.$$

Moreover, conditional on AI use, the physician follows the AI signal.

(b) For type-y patients, the physician uses AI if and only if

$$\max\left\{\underline{\alpha}, \frac{(1-\rho_y)(b+\theta\ell) + c - \theta r}{b}\right\} < \alpha_y < \frac{\rho_y b - (1-\rho_y)\theta\ell - c + \theta r}{b}.$$

Moreover, conditional on AI use, the physician follows the AI signal.

The proof of Lemma OA1 follows the same logic as the proofs of Lemmas 1 and 2 and is therefore omitted. Relative to the baseline, the ordering of type-specific AI use need not be uniform because use is jointly determined by type-dependent priors and payoff trade-offs. Nevertheless, the baseline pattern remains feasible: there are parameter values for which the physician uses AI less for type-y patients than for type-x patients. Under a disparate algorithm, integrating the physician's AI use rule in Lemma OA1 against the prior densities ($1/\overline{\alpha}$ on $(0, \overline{\alpha})$ for type-x patients and $1/(1-\underline{\alpha})$ on $(\underline{\alpha}, 1)$ for type-y patients) yields expected AI demand for each patient type. To ease notation, for type-x patients, let $\bar{d}_x(\rho_x) = \overline{\alpha} - ((1-\rho_x)b + c - \theta r)/b$ and $\tilde{d}_x(\rho_x) = ((2\rho_x - 1)b - 2c + 2\theta r)/b$. For type-$y$ patients, let $\bar{d}_y(\rho_y) = ((\rho_y b - (1-\rho_y)\theta\ell - c + \theta r)/b) - \underline{\alpha}$ and $\tilde{d}_y(\rho_y) = ((2\rho_y - 1)b - 2(1-\rho_y)\theta\ell - 2c + 2\theta r)/b$.

$$\begin{aligned} d_x^{\mathrm{D}} &= \min\{\bar{d}_x(\rho_x), \tilde{d}_x(\rho_x)\}, \\ d_y^{\mathrm{D}} &= \tilde{\beta}\min\{\bar{d}_y(\rho_y), \tilde{d}_y(\rho_y)\}. \end{aligned} \tag{OA4}$$

Note that all optima characterized in this extension lie in the region of strictly positive demand (this is verified in every numerical configuration we report), so the positive-part operator is omitted from the piecewise expressions without affecting the analysis. The AI firm then chooses (ρ_x, ρ_y) to maximize expected profit in eq. (1). Under an equal-accuracy algorithm, define $\bar{d}_x^{\mathrm{E}}(\rho) = \overline{\alpha} - ((1-\rho)b + c - \theta r)/b$, $\bar{d}_y^{\mathrm{E}}(\rho) = (\rho b - c + \theta r)/b - \underline{\alpha}$, and $\tilde{d}^{\mathrm{E}}(\rho) = ((2\rho - 1)b - 2c + 2\theta r)/b$. Then

$$d_x^{\mathrm{E}} = \min\{\bar{d}_x^{\mathrm{E}}(\rho), \tilde{d}^{\mathrm{E}}(\rho)\},$$
$$d_y^{\mathrm{E}} = \tilde{\beta}\min\{\bar{d}_y^{\mathrm{E}}(\rho), \tilde{d}^{\mathrm{E}}(\rho)\}.$$

The AI firm chooses $\rho \in (1/2, 1)$ to maximize expected profit in eq. (2).

The next proposition characterizes the resulting equilibrium accuracy choices.

Proposition OA2. *There exist parameter regions in which the AI firm develops a disparate algorithm when ℓ is below a cutoff and an equal-accuracy algorithm when ℓ exceeds that cutoff. Moreover, if $b\kappa_y > 2\tilde{\beta}\kappa_x(b + \theta\ell)$, then the associated accuracies satisfy $\rho_x^* > \rho_y^*$ whenever the disparate design is optimal. The proof characterizes the optimal accuracies under both designs.*

Proof of Proposition OA2. The proof proceeds in three steps. Step 1 characterizes the optimal accuracies under the disparate design and establishes the ranking $\rho_x^* > \rho_y^*$ under $b\kappa_y > 2\tilde{\beta}\kappa_x(b + \theta\ell)$. Step 2 characterizes the optimal accuracy under the equal-accuracy design. Step 3 establishes the liability cutoff.

Step 1: Disparate design. We begin with the disparate design. Because the physician's usage rule changes discretely at an accuracy cutoff, the AI firm's objective is piecewise in ρ_t. Conditional on supplying a disparate design, the problem separates by patient type. We therefore solve for ρ_x^* and ρ_y^* in turn. For type-x patients, define the cutoff accuracy as $\bar{\rho}_x := \frac{\overline{\alpha}b + c - \theta r}{b}$. The firm chooses $\rho_x \in (1/2, 1)$ to maximize

$$\pi_x(\rho_x) = \begin{cases} \tilde{f} \cdot \tilde{d}_x(\rho_x) - \kappa_x(\rho_x - \frac{1}{2})^2, & \text{if } \frac{1}{2} < \rho_x \le \bar{\rho}_x, \\ \tilde{f} \cdot \bar{d}_x(\rho_x) - \kappa_x(\rho_x - \frac{1}{2})^2, & \text{if } \bar{\rho}_x < \rho_x < 1. \end{cases}$$

On each region, $\pi_x(\rho_x)$ is strictly concave in ρ_x. The first-order condition yields the interior candidate $\rho_x = \frac{1}{2} + \frac{\tilde{f}}{\kappa_x}$ in the first region and $\rho_x = \frac{1}{2} + \frac{\tilde{f}}{2\kappa_x}$ in the second. The global

maximizer is obtained by checking whether the relevant candidate lies in its region; otherwise the optimum is attained at the boundary $\bar{\rho}_x$. This yields

$$\rho_x^* = \begin{cases} \frac{1}{2} + \dfrac{\tilde{f}}{\kappa_x}, & \text{if } \frac{1}{2} + \dfrac{\tilde{f}}{\kappa_x} \le \bar{\rho}_x, \\ \bar{\rho}_x, & \text{if } \frac{1}{2} + \dfrac{\tilde{f}}{2\kappa_x} \le \bar{\rho}_x < \frac{1}{2} + \dfrac{\tilde{f}}{\kappa_x}, \\ \frac{1}{2} + \dfrac{\tilde{f}}{2\kappa_x}, & \text{if } \bar{\rho}_x < \frac{1}{2} + \dfrac{\tilde{f}}{2\kappa_x}. \end{cases} \tag{OA5}$$

For type-y patients, define the corresponding cutoff as $\bar{\rho}_y := 1 - \frac{\alpha b - (c - \theta r)}{b + \theta \ell}$. The AI firm chooses $\rho_y \in (1/2, 1)$ to maximize

$$\pi_y(\rho_y; \ell) = \begin{cases} \tilde{f}\tilde{\beta} \cdot \tilde{d}_y(\rho_y) - \kappa_y(\rho_y - \frac{1}{2})^2, & \text{if } \frac{1}{2} < \rho_y \le \bar{\rho}_y, \\ \tilde{f}\tilde{\beta} \cdot \bar{d}_y(\rho_y) - \kappa_y(\rho_y - \frac{1}{2})^2, & \text{if } \bar{\rho}_y < \rho_y < 1. \end{cases}$$

Again, strict concavity holds on each region. The first-order condition yields the interior candidate $\rho_y = \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{b\kappa_y}$ in the first region and $\rho_y = \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{2b\kappa_y}$ in the second. Comparing these candidates with $\bar{\rho}_y$ gives

$$\rho_y^* = \begin{cases} \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{b\kappa_y}, & \text{if } \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{b\kappa_y} \le \bar{\rho}_y, \\ \bar{\rho}_y, & \text{if } \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{2b\kappa_y} \le \bar{\rho}_y < \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{b\kappa_y}, \\ \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{2b\kappa_y}, & \text{if } \bar{\rho}_y < \frac{1}{2} + \dfrac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{2b\kappa_y}. \end{cases} \tag{OA6}$$

Accuracy ranking. We now show that the assumption $b\kappa_y > 2\tilde{\beta}\kappa_x(b + \theta\ell)$ implies $\rho_x^* > \rho_y^*$ across all branch combinations of (OA5) and (OA6). First, (OA5) implies

$$\rho_x^* \;\ge\; \tfrac{1}{2} + \tfrac{\tilde{f}}{2\kappa_x};$$

this holds with equality on the third branch, holds on the second branch by its defining condition $\bar{\rho}_x \ge \frac{1}{2} + \frac{\tilde{f}}{2\kappa_x}$, and holds on the first branch because $\frac{\tilde{f}}{\kappa_x} > \frac{\tilde{f}}{2\kappa_x}$. Second, (OA6) implies

$$\rho_y^* \;\le\; \tfrac{1}{2} + \tfrac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{b\kappa_y};$$

this holds with equality on the first branch, holds on the second branch by its defining condition $\bar{\rho}_y < \frac{1}{2} + \frac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{b\kappa_y}$, and holds on the third branch because $\frac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{2b\kappa_y} < \frac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{b\kappa_y}$. The assumption $b\kappa_y > 2\tilde{\beta}\kappa_x(b+\theta\ell)$ is equivalent to

$$\frac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{b\kappa_y} \;<\; \frac{\tilde{f}}{2\kappa_x},$$

and therefore

$$\rho_y^* \;\le\; \tfrac{1}{2} + \tfrac{\tilde{\beta}\tilde{f}(b+\theta\ell)}{b\kappa_y} \;<\; \tfrac{1}{2} + \tfrac{\tilde{f}}{2\kappa_x} \;\le\; \rho_x^*,$$

so that $\rho_x^* > \rho_y^*$ whenever the disparate design is supplied. In particular, the disparity-liability channel, which is defined for $\rho_x > \rho_y$, is well defined at the firm's optimum.

Step 2: Equal-accuracy design. Impose $\rho_x = \rho_y \equiv \rho$. The relevant usage cutoffs are $\bar{\rho}_x$ defined in Step 1 and

$$\bar{\rho}^{(y)} := \frac{(1-\underline{\alpha})b + c - \theta r}{b},$$

so that $d_x^E = \tilde{d}^E(\rho)$ if and only if $\rho \le \bar{\rho}_x$ and $d_y^E = \tilde{\beta}\,\tilde{d}^E(\rho)$ if and only if $\rho \le \bar{\rho}^{(y)}$. Note that $\bar{\rho}^{(y)} < \bar{\rho}_x$ if and only if $\overline{\alpha} + \underline{\alpha} > 1$.

Case A: $\overline{\alpha} + \underline{\alpha} > 1$ (so $\bar{\rho}^{(y)} < \bar{\rho}_x$). The firm's objective is

$$\pi^E(\rho) = \begin{cases} \tilde{f}(1+\tilde{\beta})\,\tilde{d}^E(\rho) - (\kappa_x+\kappa_y)\big(\rho - \tfrac{1}{2}\big)^2, & \text{if } \tfrac{1}{2} < \rho \le \bar{\rho}^{(y)}, \\ \tilde{f}\big[\tilde{d}^E(\rho) + \tilde{\beta}\,\bar{d}_y^E(\rho)\big] - (\kappa_x+\kappa_y)\big(\rho - \tfrac{1}{2}\big)^2, & \text{if } \bar{\rho}^{(y)} < \rho \le \bar{\rho}_x, \\ \tilde{f}\big[\bar{d}_x^E(\rho) + \tilde{\beta}\,\bar{d}_y^E(\rho)\big] - (\kappa_x+\kappa_y)\big(\rho - \tfrac{1}{2}\big)^2, & \text{if } \bar{\rho}_x < \rho < 1. \end{cases}$$

The marginal revenues on the three regions are $2\tilde{f}(1+\tilde{\beta})$, $\tilde{f}(2+\tilde{\beta})$, and $\tilde{f}(1+\tilde{\beta})$, which are strictly decreasing across regions; combined with the strictly convex cost, $\pi^E(\rho)$ is globally concave in ρ. The region-specific first-order candidates are

$$\rho_1 := \tfrac{1}{2} + \tfrac{\tilde{f}(1+\tilde{\beta})}{\kappa_x+\kappa_y}, \qquad \rho_2 := \tfrac{1}{2} + \tfrac{\tilde{f}(2+\tilde{\beta})}{2(\kappa_x+\kappa_y)}, \qquad \rho_3 := \tfrac{1}{2} + \tfrac{\tilde{f}(1+\tilde{\beta})}{2(\kappa_x+\kappa_y)},$$

with $\rho_1 > \rho_2 > \rho_3$, and the unique maximizer is

$$\rho^* = \begin{cases} \rho_1, & \text{if } \rho_1 \le \bar{\rho}^{(y)}, \\ \bar{\rho}^{(y)}, & \text{if } \rho_2 \le \bar{\rho}^{(y)} < \rho_1, \\ \rho_2, & \text{if } \bar{\rho}^{(y)} < \rho_2 \le \bar{\rho}_x, \\ \bar{\rho}_x, & \text{if } \rho_3 \le \bar{\rho}_x < \rho_2, \\ \rho_3, & \text{if } \bar{\rho}_x < \rho_3, \end{cases} \tag{OA7}$$

truncated, if necessary, to the admissible interval $(1/2, 1)$; as in the baseline, we maintain parameter restrictions under which the optimum is interior.

Case B: $\overline{\alpha} + \underline{\alpha} \le 1$ (so $\bar{\rho}_x \le \bar{\rho}^{(y)}$). The middle region becomes $(\bar{\rho}_x, \bar{\rho}^{(y)}]$, on which $\pi^E(\rho) = \tilde{f}\big[\bar{d}_x^E(\rho) + \tilde{\beta}\,\tilde{d}^E(\rho)\big] - (\kappa_x + \kappa_y)(\rho - \frac{1}{2})^2$ with marginal revenue $\tilde{f}(1 + 2\tilde{\beta})$ and candidate $\rho_2' := \frac{1}{2} + \frac{\tilde{f}(1+2\tilde{\beta})}{2(\kappa_x+\kappa_y)}$. Since $2\tilde{f}(1+\tilde{\beta}) > \tilde{f}(1+2\tilde{\beta}) > \tilde{f}(1+\tilde{\beta})$, global concavity again holds, and ρ^* is given by

$$\rho^* = \begin{cases} \rho_1, & \text{if } \rho_1 \le \bar{\rho}_x, \\ \bar{\rho}_x, & \text{if } \rho_2' \le \bar{\rho}_x < \rho_1, \\ \rho_2', & \text{if } \bar{\rho}_x < \rho_2' \le \bar{\rho}^{(y)}, \\ \bar{\rho}^{(y)}, & \text{if } \rho_3 \le \bar{\rho}^{(y)} < \rho_2', \\ \rho_3, & \text{if } \bar{\rho}^{(y)} < \rho_3. \end{cases}$$

Let $\pi^{E*} := \pi^E(\rho^*)$. Because no term in the equal-accuracy problem involves ℓ, the maximized payoff π^{E*} is independent of ℓ.

Step 3: Existence of the liability cutoff. Let $\pi^{D*}(\ell) := \pi_x(\rho_x^*) + \pi_y(\rho_y^*; \ell)$ denote the maximized payoff under the disparate design. The type-x component does not depend on ℓ. For the type-y component, observe that for every fixed $\rho_y \in (1/2, 1)$,

$$\frac{\partial \tilde{d}_y(\rho_y)}{\partial \ell} = -\frac{2(1-\rho_y)\theta}{b} < 0, \qquad \frac{\partial \bar{d}_y(\rho_y)}{\partial \ell} = -\frac{(1-\rho_y)\theta}{b} < 0,$$

so both branches of $\pi_y(\rho_y; \cdot)$, and hence their pointwise minimum-based composition, are strictly decreasing in ℓ at every ρ_y. (We restrict attention to parameters for which type-y demand is strictly positive at the disparate optimum for all $\ell \le \widetilde{\ell}$, as holds in every

numerical configuration we report, so that $\pi^{D*}(\ell)$ is strictly decreasing on the relevant range.) Because the objective is strictly decreasing in ℓ pointwise in ρ_y, its maximum over ρ_y is strictly decreasing in ℓ, on interior and boundary branches alike; continuity of $\pi^{D*}(\ell)$ in ℓ follows from the maximum theorem. Since π^{E*} is constant in ℓ, the difference $\pi^{D*}(\ell) - \pi^{E*}$ is continuous and strictly decreasing, and therefore crosses zero at most once. Define $\widetilde{\ell} := \inf\{\ell \geq 0 : \pi^{D*}(\ell) \leq \pi^{E*}\}$ (with $\widetilde{\ell} = \infty$ if the set is empty). Then the firm strictly prefers the disparate design for $\ell < \widetilde{\ell}$ and the equal-accuracy design for $\ell > \widetilde{\ell}$; the parameter regions in the statement are those for which $\widetilde{\ell} \in (0, \infty)$, which is nonempty for the numerical configurations reported in the proof of Proposition OA3. Finally, by Step 1, whenever the disparate design is chosen the optimal accuracies satisfy $\rho_x^* > \rho_y^*$ under the assumption $b\kappa_y > 2\tilde{\beta}\kappa_x(b + \theta\ell)$[13]. *Q.E.D.*

Proposition OA3. *There exist scenarios in which, as ℓ increases, the physician uses AI for fewer type-y patients when ℓ is below a threshold, and for more type-y patients when ℓ is above that threshold. Depending on which branch of eq. (OA4) is active, the threshold is either*

$$\frac{b(\kappa_y - 2\tilde{\beta}\tilde{f})}{2\theta\tilde{\beta}\tilde{f}} \quad \textit{or} \quad \frac{b(\kappa_y - 4\tilde{\beta}\tilde{f})}{4\theta\tilde{\beta}\tilde{f}}.$$

Proof of Proposition OA3. When the relevant threshold is $\frac{b(\kappa_y - 4\tilde{\beta}\tilde{f})}{4\theta\tilde{\beta}\tilde{f}}$, the argument is identical to that in the proof of Proposition 2. We therefore focus on the case in which the threshold is $\frac{b(\kappa_y - 2\tilde{\beta}\tilde{f})}{2\theta\tilde{\beta}\tilde{f}}$. In this case, eq. (OA4) implies $d_y^{\mathrm{D}} = \tilde{\beta}\left(\frac{\rho_y^* b - (1-\rho_y^*)\theta\ell - c + \theta r}{b} - \underline{\alpha}\right)$, where $\rho_y^* = \frac{1}{2} + \frac{\tilde{\beta}\tilde{f}(b + \theta\ell)}{2b\kappa_y}$ from Proposition OA2 (Note the AI demand for type-y patients is actually $d_y^D/\overline{\alpha}$, but it suffices to analyze d_y^D as $\overline{\alpha}$ is a constant value). Differentiating with respect to ℓ and simplifying yields $\frac{\partial d_y^{\mathrm{D}}}{\partial \ell} \geq 0$ if and only if $\ell \geq \frac{b(\kappa_y - 2\tilde{\beta}\tilde{f})}{2\theta\tilde{\beta}\tilde{f}}$. Thus, holding fixed the disparate-design regime, liability reduces type-y AI use for ℓ below this cutoff and increases it above the cutoff. Non-monotonicity in equilibrium arises when the firm switches to an equal-accuracy design at a higher value of ℓ; the numerical instance in the footnote illustrates this pattern.[14] *Q.E.D.*

[13] Absent the assumption $b\kappa_y > 2\tilde{\beta}\kappa_x(b + \theta\ell)$, the accuracy ranking under the optimal disparate algorithm can reverse, that is, $\rho_x^* < \rho_y^*$. The mechanism is that without the assumption, the type-x cap $\bar{\rho}_x = (\overline{\alpha}b + c - \theta r)/b$ can bind, pushing ρ_x^* to the half-return branch $\frac{1}{2} + \frac{f}{2\kappa_x}$, whereas $\bar{\rho}_y$ cannot bind when its value exceeds one.

[14] For example, when $\underline{\alpha} = 0.4$, $\overline{\alpha} = 0.88$, $b = 0.9$, $\theta = 1$, $c = 0.36$, $r = 0.3$, $\kappa_x = 0.4$, $\kappa_y = 0.55$, $\tilde{f} = 0.28$, and $\tilde{\beta} = 0.9$, the physician uses AI for fewer type-y patients if $0 < \ell < 0.0821$ and for more type-y patients if $0.0821 < \ell < 0.4224$; the equal-accuracy design is optimal if $\ell \geq 0.4224$.

The two thresholds correspond to the cases in which d_y^{D} is determined by the first and second expressions in eq. (OA4), respectively. The intuition from our base model still applies: liability exposure trades off against the accuracy gains that higher liability induces. When liability is small, exposure concerns dominate, and the physician uses AI for fewer patients. In contrast, when liability is large, the accuracy effect dominates, and the physician uses AI for more patients.

We next examine the welfare effect of mandating equal-accuracy AI. For ease of presentation, we define the adoption cutoffs:

$$\begin{aligned}
\underline{a}_x^{\mathrm{D}} &= \frac{(1-\rho_x^*)b + c - \theta r}{b}, & \overline{a}_x^{\mathrm{D}} &= \min\left\{\overline{\alpha}, \frac{\rho_x^* b - c + \theta r}{b}\right\},\\
\underline{a}_y^{\mathrm{D}} &= \max\left\{\underline{\alpha}, \frac{(1-\rho_y^*)(b+\theta\ell) + c - \theta r}{b}\right\}, & \overline{a}_y^{\mathrm{D}} &= \frac{\rho_y^* b - (1-\rho_y^*)\theta\ell - c + \theta r}{b},\\
\underline{a}_x^{m} &= \frac{(1-\rho^m)b + c - \theta r}{b}, & \overline{a}_x^{m} &= \min\left\{\overline{\alpha}, \frac{\rho^m b - c + \theta r}{b}\right\},\\
\underline{a}_y^{m} &= \max\left\{\underline{\alpha}, \frac{(1-\rho^m)b + c - \theta r}{b}\right\}, & \overline{a}_y^{m} &= \frac{\rho^m b - c + \theta r}{b}.
\end{aligned}$$

Then patient welfare under disparate and equal-accuracy algorithms is

$$\begin{aligned}
W_x(\rho_x^*) &= \frac{1}{\overline{\alpha}}\left[\int_0^{\underline{a}_x^{\mathrm{D}}} (1-\alpha)b\,d\alpha + \int_{\underline{a}_x^{\mathrm{D}}}^{\overline{a}_x^{\mathrm{D}}} (\rho_x^* b - c)\,d\alpha + \int_{\overline{a}_x^{\mathrm{D}}}^{\overline{\alpha}} \alpha b\,d\alpha\right],\\
W_y(\rho_y^*) &= \frac{\tilde{\beta}}{\overline{\alpha}}\left[\int_{\underline{\alpha}}^{\underline{a}_y^{\mathrm{D}}} (1-\alpha)b\,d\alpha + \int_{\underline{a}_y^{\mathrm{D}}}^{\overline{a}_y^{\mathrm{D}}} (\rho_y^* b - c)\,d\alpha + \int_{\overline{a}_y^{\mathrm{D}}}^{1} \alpha b\,d\alpha\right],\\
W_x(\rho^m) &= \frac{1}{\overline{\alpha}}\left[\int_0^{\underline{a}_x^{m}} (1-\alpha)b\,d\alpha + \int_{\underline{a}_x^{m}}^{\overline{a}_x^{m}} (\rho^m b - c)\,d\alpha + \int_{\overline{a}_x^{m}}^{\overline{\alpha}} \alpha b\,d\alpha\right],\\
W_y(\rho^m) &= \frac{\tilde{\beta}}{\overline{\alpha}}\left[\int_{\underline{\alpha}}^{\underline{a}_y^{m}} (1-\alpha)b\,d\alpha + \int_{\underline{a}_y^{m}}^{\overline{a}_y^{m}} (\rho^m b - c)\,d\alpha + \int_{\overline{a}_y^{m}}^{1} \alpha b\,d\alpha\right].
\end{aligned}$$

Figure OA3 reports the welfare effects at the parameter values of the preceding footnote. The central message of the baseline persists: over a substantial region of the parameter space, mandating equal accuracy harms both patient groups. Bounded prior supports add two features. When the type-y segment is small, the mandate lowers type-x accuracy enough to curb type-x overuse, so type-x patients benefit; and when liability is high, removing the disparity-contingent exposure raises type-y use from inefficiently low levels, so type-y patients

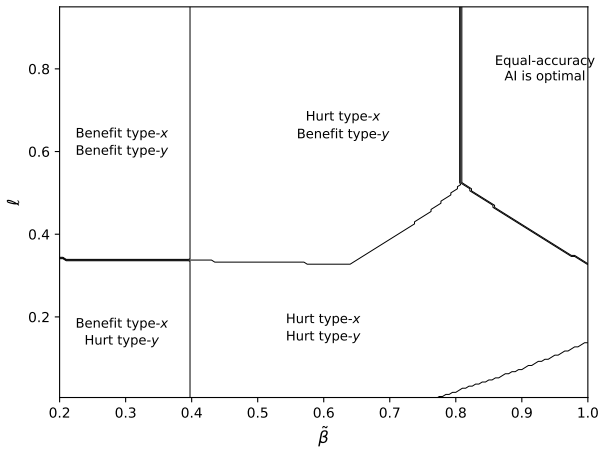


Figure OA3: Impact of mandating equal-accuracy AI on expected welfare for each patient type ($\underline{\alpha} = 0.4$, $\overline{\alpha} = 0.88$, $b = 0.9$, $\theta = 1$, $c = 0.36$, $r = 0.3$, $\kappa_x = 0.4$, $\kappa_y = 0.55$, $\tilde{f} = 0.28$).

benefit. Both groups are harmed in the intermediate region in which neither corrective force is strong.

The intuition parallels the baseline overuse logic. Because the physician's private return θr tilts use beyond the altruistic benchmark, welfare effects are governed by utilization as much as by accuracy. The mandate moderates type-x accuracy, which trims type-x overuse, and it removes the liability wedge on type-y use, which raises type-y use; each effect helps the corresponding group when the initial distortion it corrects is large, and hurts that group otherwise.