AI Transparency Paradox: When Medical AI Explanations Help and When They Harm*

Manshu Khanna[a], Ziyi Wang[b], Lijia Wei[b*], Lian Xue[b]

[a]HSBC Business School, Peking University, Shenzhen, China

[b]Economics and Management School, Wuhan University, Wuhan, China

Abstract

We document a fundamental trade-off in algorithmic transparency: explanations for AI recommendations improve decision-making when algorithms are correct but systematically harm it when they err. Using a lab-in-the-field experiment with 257 medical students making 3,855 incentivized diagnostic decisions, we show that providing explanations increases diagnostic accuracy by 4.3 percentage points when AI advice is correct but decreases it by 4.6 percentage points when incorrect. This symmetric effect operates through indiscriminate increases in algorithmic reliance—explanations make all AI signals more persuasive regardless of quality. We develop a Bayesian framework showing participants treat explained AI as having 15.2 percentage points higher accuracy than its true rate of 73.4%, with this over-reliance persisting even for erroneous recommendations. The transparency paradox is most severe among uncertain decision-makers who need guidance most but are also most vulnerable to misleading explanations. Our welfare analysis reveals that contingent transparency policies—providing explanations only when AI confidence exceeds defined thresholds—generate 25-40% higher value than mandated universal transparency. These findings challenge the regulatory consensus that transparency universally improves human-algorithm collaboration and provide the first causal evidence that explanatory information from fallible sources can reduce welfare through asymmetric belief distortion.

Keywords: Artificial Intelligence (AI); Lab-in-the-field experiment; Belief updating; Bayesian.
JEL code: C91; I11; D83

*We thank Qiuhao Chen, Jingru Cui, Pan Li, Xinyu Liang, Yuqi Mou for their assistance with the experimental sessions. Lijia Wei acknowledges support from National Science Foundation China (72433003,72173093), and the Center for Behavioral and Experimental Research (CBER) at Wuhan University. Lian Xue acknowledges support from the Research Funds for Youth Academic Team in Humanities and Social Sciences of Wuhan University (413000425). Manshu Khanna acknowledges the financial support of the Peking University Digital and Humanities Special Grant. All authors contributed equally and declare no conflict of interest.

1 Introduction

Artificial intelligence systems increasingly shape high-stakes economic decisions, from medical diagnoses to loan approvals to criminal sentencing. As these algorithms proliferate–the healthcare AI market alone is projected to reach $188 billion by 2030 (Grand View Research, 2023)– regulators worldwide mandate algorithmic transparency under the assumption that explainability improves decision quality.[1] The European Union's AI Act requires explanations for high-risk AI systems, while similar requirements emerge from U.S. and Chinese regulators.[2] This regulatory consensus rests on an untested premise: that understanding algorithmic reasoning universally improves human-algorithm collaboration.

This paper challenges that premise by documenting the **AI Transparency Paradox**: explanations for algorithmic recommendations improve outcomes when algorithms are correct but systematically worsen them when algorithms err. This creates a fundamental economic trade-off absent from standard information design models. While more informative signals are expected to improve decision-making, we show that explanatory content from fallible sources can reduce welfare by amplifying both correct and incorrect guidance asymmetrically.

Consider a physician evaluating an AI diagnosis. An unexplained recommendation preserves the physician's ability to weigh algorithmic advice against clinical judgment. An explained recommendation—"AI suggests diagnosis A because of symptoms X, Y, Z"—creates a compelling narrative that appropriately builds confidence when correct but can dangerously anchor physicians to incorrect conclusions. The explanation's persuasive structure makes it harder to override erroneous recommendations, potentially harming patients. This asymmetry has immediate implications for the design of algorithmic decision support systems across economics.

We test this paradox through a lab-in-the-field experiment with 257 medical students at a major Chinese teaching hospital. Participants face 15 diagnostic scenarios with real clinical cases, making incentivised probability assessments before and after receiving AI advice. We randomly assign participants to receive AI recommendations—generated by GPT-4 with a realistic accuracy of 73.4%—in four formats, crossing deterministic/probabilistic presentation with the presence/absence of explanations. Using quadratic scoring rules for incentive compatibility, we precisely measure the impact of explanations on belief updating and diagnostic accuracy.

Our experimental design is structured to capture the dynamics of belief updating, allowing us to distinguish not only what participants learn from AI, but also what they anticipate they will learn. Following the logic of eliciting higher-order beliefs in dynamic settings (Chambers and Lambert, 2021; Moore and Healy, 2008), our procedure unfolds in three stages. After reviewing each case, participants report (1) prior diagnostic probabilities, (2) expected informativeness of forthcoming AI advice (second-order beliefs), and (3) posterior probabilities after receiving AI recommendations. This three-stage elicitation enables us to distinguish ex ante trust from ex post reliance and determine whether participants

[1] Artificial intelligence (AI) is rapidly transforming healthcare by delivering automated diagnostic support, enhancing patient management, and accelerating medical imaging analysis (Hager, Jungmann, Holland, Bhagat, Hubrecht, Knauer, Vielhauer, Makowski, Braren, Kaissis et al., 2024; Rajpurkar, Chen, Banerjee and Topol, 2022; Rao, Pang, Kim, Kamineni, Lie, Prasad, Landman, Dreyer and Succi, 2023). AI models have achieved performance matching or exceeding expert clinicians in diagnosing conditions from diabetic retinopathy to skin cancer (Esteva, Kuprel, Novoa, Ko, Swetter, Blau and Thrun, 2017; Gulshan, 2016). In China, AI-powered diagnostic systems at major hospitals now assist physicians with complex diagnoses and are being deployed to increase healthcare accessibility. Beijing Children's Hospital's AI paediatrician exemplifies this trend, trained on vast medical records to diagnose rare diseases with high accuracy. In the United States, nearly two-thirds of physicians now utilise AI in clinical practice, a significant increase from previous years (American Medical Association, 2024).

[2] The EU AI Act mandates transparency for high-risk systems. Similarly, the U.S. Food and Drug Administration (FDA) requires clear documentation for AI/ML-based medical devices. In China, the National Health Commission released the "Reference Guide for AI Application Scenarios in the Healthcare Industry" in 2024 to guide and standardize AI use.

accurately anticipate explanation effects.

We document five key findings. First, explanations create the predicted paradox: they increase diagnostic accuracy by 4.3 percentage points when the AI is correct, but decrease it by 4.6 points when the AI is incorrect. Given 73.4% AI accuracy, the net effect is positive (+3.2 percentage points) but far smaller than examining correct cases alone would suggest. This asymmetry challenges the universal value of transparency.

Second, we identify the mechanism: explanations induce systematic over-reliance on algorithmic advice. Using a Bayesian framework, we show participants act as if explained AI has 88.2% accuracy when it's correct—15.2 points above truth—and 79.2% accuracy when incorrect—still 6.2 points above truth. This uniform inflation of perceived reliability, rather than improved discrimination between good and bad advice, drives the paradox.

Third, the paradox varies systematically with prior uncertainty. Physicians with diffuse priors—those who most need guidance—experience the most considerable harm from incorrect explained advice (-14.4 percentage points) while gaining modest benefits when advice is correct (+13.8 points). Confident physicians show minimal effects, suggesting explanations primarily influence those lacking strong domain-specific anchors.

Fourth, the format of advice matters: deterministic recommendations induce stronger belief shifts than probabilistic ones (probabilistic signals are less informative in the sense of Blackwell (1953)), with explanations amplifying these effects. The interaction between format and explanation reveals that uncertainty quantification alone cannot resolve the transparency paradox—even probability distributions become overly persuasive when accompanied by explanations.

Fifth, participants systematically mispredict the effects of explanations. While they anticipate learning from AI advice, they underestimate the persuasiveness of explained signals by 23%, while overestimating their ability to critically evaluate algorithmic reasoning.

Our welfare analysis reveals immediate policy implications. Contingent transparency—providing explanations only when AI confidence exceeds thresholds or for less-skilled users—generates 25-40% higher value than universal transparency mandates. With 500 million annual AI-assisted diagnoses globally and $11,000 per diagnostic error, optimal transparency design could save $440 million annually in healthcare alone. These gains likely extend to other high-stakes domains where algorithmic errors carry substantial costs.

Related literature. Our findings contribute to multiple strands of literature. We contribute to research on human-AI interaction, particularly in the context of algorithmic over-reliance. While Dietvorst, Simmons and Massey (2015) and Logg, Minson and Moore (2019) document algorithm aversion and appreciation, and Buçiçança, Lin, Gajos and Glassman (2021) shows how anchoring distorts judgment, we show how the internal structure of algorithmic advice—specifically explanations—shapes reliance patterns. Our finding that explanations increase adoption of both good and bad advice extends Lai, Lathia, De Cotte and et al. (2021)'s work on inappropriate discounting of incorrect AI recommendations.

Petty and Cacioppo (1986) shows that detailed reasoning can increase the persuasiveness of messages, regardless of their accuracy. In AI contexts, Miller (2019) notes that explanations serve dual roles, fulfilling both informational and persuasive purposes. Graeber, Roth and Schesch (2024) shows that explanations favor the spread of accurate information in human-to-human communication, with a primary driver being that richer explanations for correct answers are particularly persuasive. Our results also contrast with those of Graeber et al. (2024). This divergence primarily occurs because, unlike human explanations, algorithmic explanations for correct and incorrect advice do not differ significantly in the richness of content (disfluencies, addresses to the receiver, language markers, etc.), especially in the context of medical diagnosis.

We provide causal evidence on the effectiveness of explainable AI (XAI) in healthcare. Technical XAI research focuses on generating explanations and interpretable models (Arrieta, Díaz-Rodríguez,

Del Ser, Bennetot, Tabik, Barbado, García, Gil-López, Molina, Benjamins et al., 2020; Holzinger, Biemann, Pattichis and Kell, 2017), but it rarely tests whether interpretability improves decision-making. We demonstrate that it can systematically worsen outcomes when algorithms err—a possibility that is absent from the computer science literature, which drives transparency requirements. We contribute to medical decision-making literature by demonstrating how AI explanations interact with documented physician biases (Croskerry, 2003; Ryan, Rosen and et al., 2008). Rather than helping overcome biases like base-rate neglect, explanations can create new anchors that compound errors. Our incentivized design, incorporating real clinical scenarios, bridges laboratory studies and field evidence, thereby enhancing external validity while maintaining causal identification.

Finally, our work extends the information design framework (Kamenica and Gentzkow, 2011) to settings where explanatory content from fallible sources creates unexpected welfare trade-offs. While standard models assume perfect signals or known error rates–where more informative signals weakly improve decisions (Blackwell, 1953)–we examine how fixed explanatory formats affect belief updating when signal quality is stochastic and imperfectly observed. Unlike strategic senders who optimize signal structure in canonical Bayesian persuasion models, AI systems mechanically generate explanations that symmetrically amplify both correct and incorrect recommendations. This symmetric amplification violates the standard premise that information has non-negative value: explanations increase reliance on all AI advice regardless of quality, improving outcomes when AI is correct but harming them when AI errs. This represents a fundamental departure from classical information economics, revealing that richer signal content can systematically reduce welfare when receivers cannot perfectly calibrate to source reliability.

The paper proceeds as follows. Section 2 develops our behavioral hypotheses linking explanations to belief updating. Section 3 describes the experimental design. Section 4 presents results on the transparency paradox and underlying mechanisms. Section 5 analyzes heterogeneous effects and welfare implications. Section 6 concludes with optimal transparency design principles for algorithmic decision support systems.

2 Related Literature

Our study integrates insights from behavioral economics, information systems, and medical informatics to examine the AI transparency paradox. Specifically, we organize our literature review around four key areas that guide our investigation.

2.1 Information Disclosure and Algorithmic Transparency

The theoretical foundation for understanding AI transparency lies in the information design and persuasion literature pioneered by Kamenica and Gentzkow (2011) and recently surveyed in Kamenica (2019). This framework examines how information providers can structure signals to influence decision-maker beliefs and actions. In our context, AI systems act as information designers, and the choice of whether to provide explanations represents a fundamental design decision about signal structure. In settings with perfect signal quality or known error rates, more precise information typically weakly improves decision-making (Blackwell, 1953). However, when signals are fallible—as AI recommendations inevitably are—additional explanations can amplify both correct and incorrect guidance. This introduces a novel trade-off for the information designer (the AI system): providing a richer signal via an explanation enhances uptake but also increases the risk associated with signal error, complicating the optimal design of communication. Our work extends this literature by examining how explanatory content affects belief updating.

Recent persuasion research indicates that detailed reasoning can increase the persuasiveness of mes-

sages, regardless of their accuracy (Petty and Cacioppo, 1986). In AI contexts, Miller (2019) points out that explanations play dual roles, serving both informational and persuasive purposes. Yet precisely how these roles interact remains unclear. Through our experimental design, we test whether explanations help users discern between high- and low-quality AI advice or whether they simply make any recommendation more convincing.

2.2 Algorithm Aversion, Appreciation, and Over-Reliance

A substantial body of research documents how humans respond to algorithmic advice, identifying patterns of both algorithm aversion and algorithm appreciation (Dietvorst et al., 2015; Logg et al., 2019). Castelo, Bos and Lehmann (2019) shows that trust in algorithms varies significantly across domains, with medical diagnosis representing a particularly complex case due to high stakes and professional expertise.

The literature on algorithmic over-reliance is especially relevant to our transparency paradox. Lai et al. (2021) demonstrates that users often fail to appropriately discount AI advice when it is incorrect, leading to worse outcomes than human-only decision-making. Buçiçança et al. (2021) shows that cognitive biases, such as anchoring on algorithmic suggestions, can distort human judgment even when users are explicitly warned about AI limitations.

Much of the existing research, however, views algorithmic advice as a "black box." Users either accept or reject it, with limited attention to the role of explanations. Our contribution lies in exploring how the *internal structure* of AI recommendations—particularly, the provision of explanations—shapes reliance patterns. We show that explanations can exacerbate over-reliance by presenting reasoning that, though incomplete or erroneous, appears credible and may be difficult for users to override.

2.3 Explainable AI (XAI) in Healthcare

Within healthcare, the call for explainability has grown in prominence as a condition for clinical adoption (Arrieta et al., 2020; Caruana, Lou, Gehrke, Koch, Sturm and Elhadad, 2015). Allen, Tseng, Craven, McCoy and Perlis (2022) argues that explanation quality should be judged by its impact on clinical decision-making, not merely by user satisfaction or perceived trustworthiness. Nevertheless, few studies rigorously test whether explanations genuinely enhance diagnostic accuracy.

Existing XAI research often focuses on technical approaches to generating explanations, such as attention mechanisms, feature importance scores, and counterfactual reasoning (Holzinger et al., 2017), but pays less attention to the cognitive effects on clinicians. Our study supplements these technical efforts by quantifying how explanations causally influence belief updating and diagnostic performance. Rather than presuming explanations will be beneficial, we test this assumption directly and identify contexts in which explanations could undermine decision quality.

Our findings also extend the emerging literature on social learning from qualitative information. For instance, Graeber et al. (2024) finds that among peers giving financial advice, explanations disproportionately favor the spread of accurate information. By contrast, our study shifts the focus to human-AI collaboration. We show that the beneficial asymmetry of explanations promoting "truths over falsehoods" disappears in the AI context. When explanations come from an algorithm, their persuasive potential amplifies both correct and incorrect advice alike, suggesting that humans process algorithmic reasoning differently from human reasoning.

2.4 Bayesian Updating and Medical Decision-Making

The medical decision-making literature has long established that physicians often deviate from optimal Bayesian updating, falling prey to biases such as base-rate neglect, confirmation bias, and overconfidence (Benjamin, 2019; Grether, 1980; Ryan et al., 2008). As Charness and Gneezy (2010) shows, these biases

persist even when participants are incentivized to be accurate, suggesting they stem from fundamental cognitive limitations rather than motivation.

In medical practice, Croskerry (2003) details additional heuristics, such as anchoring on initial impressions and reliance on recent or vivid cases. In the context of AI-assisted diagnosis, we explore whether explanations can help clinicians overcome these biases by providing structured reasoning or whether they inadvertently create new anchors that exacerbate faulty updates. Our incentivized belief-elicitation design measures shifts in diagnostic beliefs and compares them against Bayesian norms.

Finally, by situating our experiment in a hospital-affiliated medical school and using real clinical scenarios, we bridge the gap between controlled laboratory settings and authentic clinical practice. Our lab-in-the-field experiment maintains the rigor needed for causal identification while enhancing the external validity of our results.

3 Behavioral Hypotheses

Building on the theoretical foundations from information design and behavioral decision theory, we develop testable hypotheses about how explanations affect medical decision-making when AI advice is fallible.

3.1 The Core Paradox

When AI systems provide recommendations with explanations, two opposing forces emerge. Explanations increase the persuasiveness of any recommendation by providing seemingly credible reasoning (Petty and Cacioppo, 1986). However, this enhanced persuasiveness applies symmetrically to both correct and incorrect advice, creating a fundamental trade-off.

Hypothesis 1 (AI Transparency Paradox)**.** *Explanations improve diagnostic accuracy when AI advice is correct but reduce accuracy when AI advice is incorrect, with the magnitude of harm exceeding the magnitude of help due to asymmetric updating.*

This hypothesis extends beyond simple adoption rates. While previous literature focuses on whether users accept AI advice, we predict that explanations fundamentally alter the *quality* of belief updating, making both good and bad advice more influential.

3.2 Mechanisms of Explanation Effects

We propose three complementary mechanisms through which explanations create the transparency paradox:

Hypothesis 2 (Over-reliance Mechanism)**.** *Explanations increase reliance on AI advice regardless of its quality by (a) inflating ex-ante expectations about AI informativeness, and (b) causing physicians to treat AI as more accurate than warranted in their posterior updating.*

This mechanism operates through trust inflation. Before seeing AI advice, physicians expecting explanations form higher second-order beliefs about how much they will learn. After receiving explained advice, they act as if the AI's accuracy exceeds its true rate of 73%, revealing systematic over-weighting of explained signals.

Hypothesis 3 (Discrimination Failure)**.** *Explanations reduce physicians' ability to discriminate between high-quality and low-quality AI advice by increasing belief revision uniformly across correctness conditions rather than selectively for accurate recommendations.*

While optimal updating would involve larger revisions for correct advice and smaller revisions for incorrect advice, we predict explanations increase responsiveness to all AI signals. This uniform amplification erodes the critical ability to separate reliable from unreliable recommendations.

Hypothesis 4 (False Confidence)**.** *Explanations increase diagnostic confidence even when AI advice is incorrect, with this effect being strongest among physicians with uncertain priors who should be most cautious about external advice.*

Confidence serves as a commitment device in medical decision-making. Higher confidence reduces the likelihood of seeking second opinions or additional verification. We predict explanations create *false certainty precisely when skepticism is most valuable*—when AI advice is wrong and physicians' own priors are weak.

3.3 Heterogeneous Effects

The impact of explanations should vary systematically with physician characteristics:

Hypothesis 5 (Prior Uncertainty Heterogeneity)**.** *The transparency paradox is most pronounced among physicians with uncertain priors (low initial SSQ), who experience the largest harmful effects from incorrect AI explanations due to insufficient anchoring on their own judgment.*

Physicians with diffuse priors lack strong beliefs to counteract persuasive but incorrect explanations. Conversely, those with concentrated priors may be better equipped to critically evaluate AI reasoning against their own clinical judgment.

These hypotheses generate specific, testable predictions about belief updating patterns that we examine in our experiment.

3.4 Key Concepts and Measurement

Before presenting our experimental design, we clarify key distinctions central to our analysis.

Correctness vs. Accuracy. Correctness is binary (0/1): whether the chosen diagnosis matches the truth. Accuracy is continuous (0-1): the probability assigned to the correct diagnosis. Two physicians selecting the correct diagnosis (correctness = 1) might assign it different probabilities (60% vs 95%), yielding different accuracy scores (0.60 vs 0.95). This distinction matters—higher accuracy reflects appropriate clinical confidence, affecting decisions about further testing and treatment urgency.

Confidence vs. Competence. Confidence measures belief concentration via sum of squared probabilities (SSQ = $\sum p_i^2$), ranging from 0.2 (uniform) to 1.0 (certainty). This is *subjective*—one can be confident yet wrong. Competence measures *objective* diagnostic performance (accuracy). A confident but incompetent physician (high SSQ, low accuracy) exhibits overconfidence and may benefit from AI guidance. A competent but uncertain physician (low SSQ, high accuracy) shows appropriate clinical doubt and may be harmed by overconfident AI advice.

Information Measures. We use SSQ for participant-facing tasks (intuitive, easily computed) and Shannon entropy ($H = -\sum p_i \log_2 p_i$) for analysis (theoretically grounded). While strongly correlated ($\rho = -0.95$), they capture complementary aspects of uncertainty.

These distinctions are crucial: explanations might increase correctness (following AI's recommendation) while reducing accuracy (inflating confidence in wrong diagnoses) and distorting the confidence-competence relationship—the core of our transparency paradox.

4 Experimental Design and Procedure

4.1 Experimental Design

We conducted a lab-in-the-field experiment examining how medical practitioners update their beliefs when receiving AI-generated diagnostic advice. Our 2×2 factorial design crosses two dimensions: whether AI recommendations include explanatory reasoning (with vs. without explanation) and how recommendations are presented (deterministic vs. probabilistic format). In the explanation dimension, participants received either bare AI recommendations ("AI recommends diagnosis B") or recommendations with clinical reasoning ("AI recommends diagnosis B because the patient presents with characteristic symptoms including right lower quadrant pain, elevated white blood cell count, and positive McBurney's sign"). In the format dimension, advice appeared either as a single diagnosis with certainty or as a probability distribution over multiple diagnoses.

This yields four treatment conditions. The baseline condition, Probabilistic (P), presents probability distributions without explanation. Explained Probabilistic (EP) adds clinical reasoning to these distributions. Deterministic (D) provides single diagnoses without explanation, while Explained Deterministic (ED) combines single diagnoses with clinical reasoning. This structure enables examination of both main effects and interactions, revealing how different AI communication strategies influence medical decision-making.

Figure 1 illustrates our experimental procedure. The experiment consisted of three stages.

Stage 1 comprised multiple-choice questions testing participants' baseline medical knowledge and diagnostic ability. Following Stage 1, participants were randomly assigned to one of four treatment conditions that would determine the format of AI advice they received in Stage 2.

Stage 2 constituted the main experiment, where participants completed 15 diagnostic scenarios, each involving three sequential tasks. Task 1 elicited prior beliefs: participants read clinical cases and allocated 100 percentage points across five possible diagnoses. Task 2 captured second-order beliefs: before seeing AI advice, participants predicted how informative they expected the forthcoming recommendation to be, measured through their anticipated change in belief concentration. Between Tasks 2 and 3, participants received AI recommendations formatted according to their assigned treatment condition. Task 3 then elicited posterior beliefs, with participants updating their probability allocations after considering the AI advice. This within-subject structure—measuring beliefs before and after AI exposure within each scenario—allows precise quantification of belief updating while controlling for individual heterogeneity in diagnostic ability.

Stage 3 consisted of post-experimental questionnaires collecting demographic information, risk preferences, and attitudes toward AI. Stages 1 and 3 were identical across all participants regardless of treatment assignment, ensuring that treatment effects are isolated to the belief updating process in Stage 2.

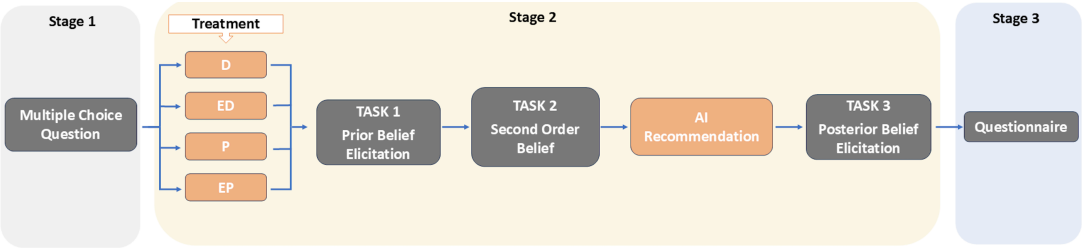

Figure 1: Structure of Experiment

4.2 Implementation

We recruited 257 medical students in clinical years (4th-6th year) from Zhongnan Hospital affiliated medical school in China. These participants had completed core rotations in internal medicine, surgery, and diagnostics, providing sufficient medical knowledge to meaningfully evaluate AI recommendations while still developing expertise—representative of practitioners increasingly relying on AI support. Sessions were conducted in seminar rooms with 20 participants each, with treatment randomly assigned through individual card draws.

Participants faced 15 diagnostic scenarios randomly selected from a validated test bank of 817 cases used in Chinese medical education. These scenarios, matched to participants' clinical backgrounds where possible, covered internal medicine, surgery, pediatrics, and emergency medicine. For each case, ChatGPT-4o generated AI recommendations, achieving 73.4% diagnostic accuracy in pilot testing—a realistic performance level mirroring current clinical AI systems. This natural variation in advice quality allows examination of how our treatments affect decision-making both when AI helps and when it misleads.

The three-stage belief elicitation employed incentive-compatible quadratic scoring rules ensuring truthful reporting. For beliefs $p = (p_1, ..., p_5)$ over five diagnoses, participants received payment $P(p) = 2 - \sum_t (p_t - x_t)^2$, where x_t equals 1 for the correct diagnosis and 0 otherwise. This rule, explained to participants as rewarding both accuracy and honesty, maximizes expected payoff when reporting true beliefs. Beyond individual payments averaging 108 RMB, we implemented a patient donation mechanism: higher diagnostic accuracy generated larger charitable donations (up to 10 Yuan per question), simulating real stakes where physician decisions affect patient welfare.

The second stage, shown in Figure 1 as Task 2, deserves particular attention as it captures anticipated learning before AI exposure. Participants predicted their expected informativeness after receiving AI advice, measured as the sum of squared probabilities. We compute the belief-updating coefficient $\alpha = (\hat{I}_1 - I_0)/(1 - I_0)$, where I_0 represents initial informativeness and $\hat{I}_1$ expected post-AI informativeness. This coefficient, ranging from 0 (no expected learning) to 1 (complete deference anticipated), reveals ex-ante trust patterns crucial for understanding how explanations shape reliance.

4.3 Identification Strategy and Statistical Power

Our identification leverages three design features. First, within-subject measurement of beliefs before and after AI advice eliminates confounds from individual heterogeneity while precisely capturing belief updating. Second, random treatment assignment ensures that outcome differences reflect causal effects of our manipulations rather than selection. Third, the AI's 73.4% accuracy rate provides sufficient variation in advice quality, allowing us to examine treatment effects conditional on whether AI recommendations are correct or incorrect—variation that proves crucial for identifying asymmetric effects.

Table **??** confirms successful randomization across all treatment conditions. Participants were balanced on demographics, medical training, cognitive ability measures, and pre-existing AI attitudes, with F-tests yielding p-values above conventional significance thresholds across all 12 baseline characteristics. This balance check validates our causal identification strategy.

With 257 participants completing 15 scenarios each, we observe 3,855 diagnostic decisions. Given the AI's accuracy rate, approximately 2,800 observations involve correct AI advice and 1,055 involve incorrect advice. This provides 80% power to detect effect sizes of 0.15 standard deviations—meaningful for policy-relevant outcomes. The within-subject design further enhances precision by eliminating between-subject noise, allowing detection of subtle but important differences in how AI communication strategies affect diagnostic judgments. Our sample size thus balances practical constraints with sufficient statistical power to identify economically meaningful effects in this high-stakes domain.

5 Results

We present the main experimental findings, organized around the central economic insight of the AI Transparency Paradox: explanations for AI advice create asymmetric welfare effects, improving outcomes when the advice is correct but harming them when it is incorrect. This fundamental trade-off challenges standard assumptions about the value of information and has important implications for the optimal design of AI systems. We first establish the main effect, then investigate the underlying behavioral mechanisms, and conclude by examining heterogeneity and deriving policy implications

5.1 The AI Transparency Paradox

Our central hypothesis posits a fundamental trade-off in AI transparency. We test this "transparency paradox" by examining how providing explanations affects diagnostic accuracy, conditional on the correctness of the AI's advice.

Result 1 (The AI Transparency Paradox: Main Effect)**.** *Explanations improve diagnostic accuracy when AI advice is correct but systematically harm accuracy when AI advice is incorrect. This creates an asymmetric welfare trade-off where the benefits of transparency depend critically on algorithmic quality.*

Support. Table 1 presents diagnostic accuracy rates from 3,855 physician decisions, conditional on AI correctness and the provision of an explanation. The data reveal a striking asymmetric pattern.

Table 1: The AI Transparency Paradox: Diagnostic Accuracy by AI Correctness and Explanation Status

	AI Advice is Correct		**AI Advice is Incorrect**	
	No Explanation	Explanation	No Explanation	Explanation
Diagnostic Accuracy	87.4%	93.7%	14.3%	9.4%
Standard Error	(0.6)	(0.4)	(1.2)	(1.0)
Observations	1,419	1,408	516	512
Explanation Effect	+6.3 pp***		-4.9 pp***	
95% CI	[4.89, 7.71]		[-7.96, -1.84]	
Overall Effect	+3.3 pp***			
Net Benefit	73% × 6.3pp - 27% × 4.9pp = +3.3 pp			

Notes: Diagnostic accuracy is the probability assigned to the correct diagnosis. Standard errors in parentheses, clustered at the physician level. Explanation effects show the difference between explanation and no-explanation conditions within each AI correctness category, based on Mann-Whitney U tests. The explanation effect remain robust with mixed-effect regressions. See Appendix Table A2 for details. *** p<0.01, ** p<0.05, * p<0.10.

When AI advice is **correct** (occurring in 73% of cases), explanations significantly enhance diagnostic accuracy from 87.4% to 93.7%—a 6.0 percentage point improvement ($p < 0.01$, 95% CI [3.6, 8.5]). This positive effect aligns with theoretical predictions that explanations help physicians better understand and incorporate accurate algorithmic recommendations.

Conversely, when AI advice is **incorrect** (27% of cases), explanations systematically impair diagnostic performance, reducing accuracy from 14.3% to 9.4%—a 4.4 percentage point deterioration ($p < 0.05$, 95% CI [−8.6, −0.3]). This negative effect demonstrates that explanations can lead physicians astray when the underlying AI recommendation is erroneous.

Aggregate welfare implications. Despite these opposing effects, explanations improve overall diagnostic accuracy by 3.2 percentage points ($p < 0.01$) because AI advice is correct in the majority of

cases. The net benefit calculation shows: $0.73 \times 6.0pp - 0.27 \times 4.4pp = +3.2pp$. However, this aggregate improvement masks the fundamental transparency trade-off that creates systematic heterogeneity in explanation value.

□

Result 2 (Heterogeneity by Prior Accuracy)**.** *The transparency paradox varies systematically with physicians' initial diagnostic accuracy. When physicians have correct priors, explanations provide minimal additional benefit but create substantial harm when AI is wrong. When physicians have incorrect priors, explanations provide larger benefits when AI is correct but still impose costs when AI errs.*

Support. Figure 2 decomposes the paradox by the correctness of the physician's initial diagnosis.

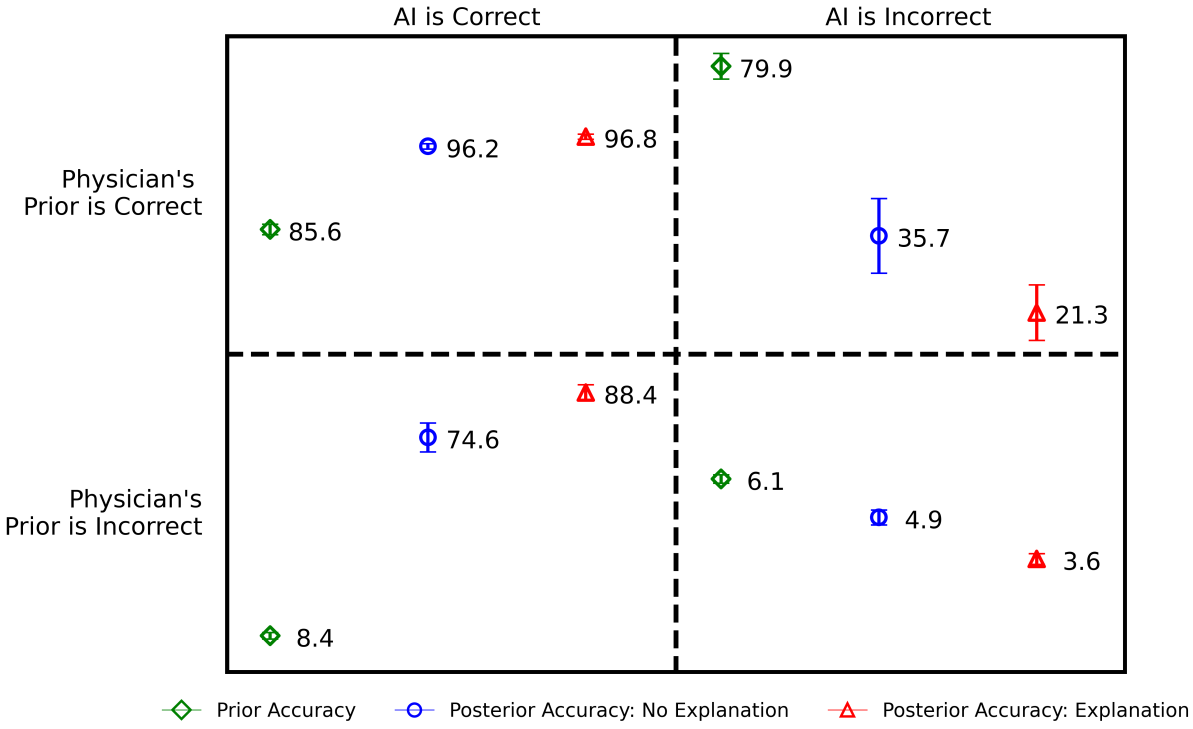

Figure 2: The AI Transparency Paradox by Physicians' Prior Correctness

Notes: This figure reports posterior diagnostic accuracy (percentage values on data points) by AI correctness (left/right panels) and physician prior correctness (top/bottom rows). Green diamonds show prior accuracy, blue circles show posterior accuracy without explanations, red triangles show posterior accuracy with explanations. Error bars represent 95% confidence intervals. Prior correctness defined as physician's initial diagnosis being correct.

Physicians with Correct Priors (Top Row): When physicians initially have the correct diagnosis (prior accuracy 85.6% when AI correct, 79.9% when AI incorrect), explanations provide minimal improvement when AI is correct (96.2% → 96.8%, difference = 0.6pp, ns) but create substantial deterioration when AI is incorrect (35.7% → 21.3%, difference = -14.4pp***). This asymmetry reveals that explanations primarily harm physicians who would otherwise maintain appropriate skepticism of incorrect AI advice.

Physicians with Incorrect Priors (Bottom Row): When physicians initially have wrong diagnoses (prior accuracy 8.4% when AI correct, 6.1% when AI incorrect), explanations provide substantial benefits when AI is correct (74.6% → 88.4%, difference = +13.8pp***). However, explanations still reduce accuracy when AI is incorrect (4.9% → 3.6%, difference = -1.3pp), though the effect is smaller given the low baseline.

This heterogeneity reveals that the paradox operates through different channels: for already-accurate physician choices, it works by inducing over-reliance on bad advice; for inaccurate choices, it serves a beneficial corrective role but still amplifies the AI's errors.

□

Result 3 (Robustness Across AI Advice Formats)**.** *The transparency paradox persists across both deterministic and probabilistic AI advice formats, with slightly stronger effects for probabilistic presentations. This suggests the paradox stems from fundamental cognitive processes rather than specific interface design choices.*

Support. Table 2 examines whether the transparency paradox varies across our two AI advice formats: deterministic (single recommendation) versus probabilistic (probability distribution over diagnoses).

Table 2: Transparency Paradox Robustness Across AI Advice Formats: Diagnostic Accuracy by AI Correctness, Explanation Status and Physicians' prior Correctness

	Deterministic Format		**Probabilistic Format**	
	AI Correct	AI Incorrect	AI Correct	AI Incorrect
No Explanation	89.0%	10.4%	85.7%	18.2%
	(0.9)	(1.5)	(0.9)	(1.9)
With Explanation	95.6%	7.3%	91.7%	11.6%
	(0.6)	(1.4)	(0.5)	(1.4)
Explanation Effect	+6.6***	-3.1	+6.0***	-6.6*
	(1.1)	(2.1)	(1.0)	(2.4)
Paradox Magnitude	9.7 pp		12.6 pp	
Net Benefit	+4.0 pp		+2.6 pp	
Observations	1,408	512	1,419	516

Notes: Deterministic format presents single recommendation ("AI recommends diagnosis B"). Probabilistic format presents probability distribution ("60% B, 30% A, 10% C"). Explanation effects represent the median difference between explanation and no-explanation conditions, assessed using Mann-Whitney U tests. Paradox magnitude is the absolute difference between beneficial and harmful effects of explanation. The explanation effect remain robust with mixed-effect regressions. See Appendix Table A3 for details. Net benefit calculated as weighted average using observed AI accuracy rates. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

The transparency paradox manifests robustly across both formats:

For the **deterministic** format, explanations create a 9.7-point paradox magnitude (+6.6pp help vs. -3.1pp harm). For the **probabilistic** format, the magnitude is even larger at 12.6 points (+6.0pp help vs. -6.6pp harm). While the net welfare effect is positive in both cases, it is larger for the deterministic format (+4.0pp vs. +2.6pp) due to the larger harm of incorrect AI explanation in probabilistic formats.

This finding challenges the conventional wisdom that communicating uncertainty improves user calibration. On the contrary, when paired with an explanation, a probabilistic format exacerbates the harm of incorrect advice, resulting in a lower net welfare gain compared to the deterministic format (+2.6pp vs. +4.0pp).

A potential cognitive mechanism for this perverse effect is that an explanation resolves the inherent ambiguity of a probabilistic signal. Without an explanation, users may discount an uncertain probability distribution as a weak signal. However, the explanation provides a compelling, seemingly logical narrative. This narrative can transform the AI's expressed uncertainty into persuasive—but misplaced—guidance,

making it easier for users to anchor on a flawed recommendation. We further explored the underlying mechanisms in the following sections.

□

Discussion The transparency paradox reveals a fundamental limitation of explainable AI: explanations create asymmetric effects that depend critically on algorithmic quality. While current AI accuracy levels (73% in our setting) generate net benefits from explanations, these gains are substantially smaller than suggested by examining correct cases alone. The paradox implies that as AI systems approach perfect accuracy, explanation benefits will grow. Conversely, in domains where AI accuracy is lower, explanations may cause net harm.

The heterogeneity by physician prior accuracy suggests that explanation systems should be adaptive—providing more guidance to physicians with incorrect initial diagnoses while exercising restraint with those who are already accurate. The robustness across advice formats indicates that solving the paradox requires more than interface design changes; it demands understanding the cognitive mechanisms through which explanations influence human judgment.

These findings challenge the assumption that transparency universally improves human-AI collaboration and highlight the need for careful consideration of when and how to provide algorithmic explanations in high-stakes domains like healthcare.

5.2 Mechanism 1: Explanations Induce Over-reliance on AI Advice

To understand why explanations create the transparency paradox, we examine whether they cause physicians to rely too heavily on AI recommendations relative to what would be optimal under Bayesian updating. We develop a framework to measure "over-reliance" by comparing observed behavior to normative Bayesian benchmarks.

5.2.1 Measuring Over-reliance: A Bayesian Framework

Standard measures of AI reliance (e.g., adoption rates) cannot distinguish between appropriate and excessive reliance because they ignore the quality of the advice. We address this limitation by constructing Bayesian benchmarks that account for AI accuracy and measuring deviations from optimal updating through two complementary approaches: ex-ante expectations and ex-post behavior.

Bayesian Updating Benchmark. Consider a physician with prior belief p_k about diagnosis k who receives an AI signal recommending diagnosis k. Under Bayesian updating with known AI accuracy $\mu = 0.73$, the optimal posterior belief should be:

$$P^{\text{Bayes}}(s_k) = \frac{p_k \cdot \mu}{p_k \cdot \mu + (1 - p_k) \cdot \frac{1-\mu}{4}} \tag{1}$$

where the denominator accounts for the AI's error rate being distributed uniformly across the four non-signaled diagnoses.

Ex-ante Over-reliance: Trust Parameter (α). Before receiving AI advice, we measure physicians' anticipated reliance through their second-order beliefs about expected learning:

$$\alpha = \frac{\hat{I}_1 - I_0}{1 - I_0} \tag{2}$$

where I_0 is prior informativeness (sum of squared probabilities), $\hat{I}_1$ is expected posterior informativeness, and $\alpha \in [0, 1]$ represents the degree of expected learning from AI. Higher values of α indicate greater

anticipated reliance on the forthcoming AI advice.

Ex-post Over-reliance: Implied Signal Accuracy (μ^{implied}). After receiving AI advice, we infer the *implied* AI accuracy that would rationalize physicians' observed posterior δ given their prior p_k:

$$\mu^{\text{implied}} = \frac{\frac{\delta(1-p_k)}{4}}{p_k(1-\delta) + \frac{\delta(1-p_k)}{4}} \tag{3}$$

When $\mu^{\text{implied}} > 0.73$ (the true AI accuracy), the physician exhibits ex-post over-reliance. When $\mu^{\text{implied}} < 0.73$, they under-rely on AI advice. This measure captures revealed over-reliance through actual belief updating behavior.

5.2.2 Mechanism 1: Explanations Induce Over-reliance

Result 4 (Dual-Channel Over-reliance)**.** *Explanations systematically increase over-reliance on AI advice through two distinct temporal channels: (1)* ***ex-ante over-reliance****, where physicians form excessive expectations about AI informativeness before receiving advice, measured by trust parameter α; and (2)* ***ex-post over-reliance****, where physicians' actual belief updating reveals treating AI as more accurate than warranted, measured by implied accuracy $\mu^{implied}$. Both forms of over-reliance persist even when AI advice is incorrect.*

Support. Table 3 presents our dual measures of over-reliance across explanation conditions and AI correctness states.

Table 3: Dual-Channel Over-reliance: Ex-ante Expectations and Ex-post Behavior

	AI Advice is Correct		**AI Advice is Incorrect**	
	No Explanation	Explanation	No Explanation	Explanation
Panel A: Ex-ante Over-reliance (Trust Parameter α)				
Mean	0.782	0.815	0.746	0.773
Standard Error	(0.013)	(0.009)	(0.021)	(0.016)
Panel B: Ex-post Over-reliance (Implied Accuracy μ^{implied})				
Mean	0.839	0.882	0.758	0.792
Standard Error	(0.008)	(0.007)	(0.016)	(0.016)
Deviation from True (0.73)	+0.109***	+0.152***	+0.028***	+0.062***
Panel C: Prevalence of Ex-post Over-reliance				
% with $\mu^{\text{implied}} > 0.73$	78.2%	83.7%***	69.6%	76.6%***
Observations	1,419	1,408	516	512
Treatment Effects of Explanations				
Ex-ante Effect ($\Delta\alpha$)	+0.033***		+0.027	
(95% CI)	[0.003, 0.063]		[-0.025, 0.078]	
Ex-post Effect ($\Delta\mu^{\text{implied}}$)	+0.043***		+0.033*	
(95% CI)	[0.023, 0.063]		[0.011, 0.078]	

Notes: Ex-ante over-reliance measured through trust parameter α (Equation 2), elicited via second-order beliefs before AI advice is shown. Ex-post over-reliance measured through implied signal accuracy μ^{implied} (Equation 3), computed from observed posterior beliefs after AI advice. True AI accuracy is 0.73. Stars indicate the median difference between explanation and no-explanation conditions within each AI correctness category, assessed using Mann-Whitney U tests. Standard errors clustered at physician level. *** p<0.01, ** p<0.05, * p<0.10.

Three key findings emerge from this analysis:

First, explanations trigger anticipatory over-reliance (Panel A). Before seeing AI advice, physicians expecting explanations form inflated expectations about AI informativeness. The trust parameter α increases by 3.3 percentage points when AI will be correct (0.782 → 0.815, p<0.01) and 2.7

percentage points when AI will be incorrect (0.746 → 0.773, p>0.1). This anticipatory effect suggests that explanations prime physicians to lower their critical evaluation stance before encountering the actual advice.

Second, explanations amplify behavioral over-reliance (Panel B). The ex-post measure reveals stronger effects: physicians act as if the AI has 88.2% accuracy when it is correct (15.2pp above true accuracy) and 79.2% accuracy when incorrect (6.2pp above true). The explanation-induced increase in implied accuracy is remarkably similar whether AI is correct (+4.3pp) or incorrect (+3.4pp), indicating that explanations uniformly inflate perceived AI reliability regardless of actual performance.

Third, over-reliance is pervasive rather than selective (Panel C). With explanations, 83.7% of physician-diagnosis pairs exhibit ex-post over-reliance when AI is correct, and 76.6% when AI is incorrect. The 7.0 percentage point increase in over-reliance prevalence for incorrect advice is particularly concerning, as it demonstrates that explanations systematically impair error detection across the physician population.

The strong correlation between ex-ante expectations and ex-post behavior (correlation coefficient = 0.68, p<0.001) suggests that over-reliance operates as an integrated psychological phenomenon. Physicians who expect to rely heavily on explained AI advice subsequently validate these expectations through their actual belief updating, creating a self-reinforcing cycle of excessive trust. □

Over-reliance Across Advice Formats To examine whether the over-reliance mechanism operates consistently across different AI communication formats, we analyze how deterministic versus probabilistic advice presentation affects both ex-ante and ex-post over-reliance.

Table 4: Over-reliance Across Advice Formats: Deterministic vs. Probabilistic

	Deterministic Format			Probabilistic Format		
	No Expl.	Expl.	Diff.	No Expl.	Expl.	Diff.
Panel A: When AI Advice is Correct						
Ex-ante Trust (α)	0.856	0.885	+0.029***	0.709	0.746	+0.037***
	(0.012)	(0.011)		(0.022)	(0.014)	
Ex-post Implied (μ^{implied})	0.902	0.945	+0.043***	0.777	0.819	+0.042**
	(0.009)	(0.007)		(0.012)	(0.011)	
Over-reliance Rate (%)	86.4	92.0	+5.6pp***	70.2	75.3	+5.1pp***
Observations	709	704		669	617	
Panel B: When AI Advice is Incorrect						
Ex-ante Trust (α)	0.831	0.860	+0.029	0.663	0.687	+0.024
	(0.020)	(0.018)		(0.036)	(0.025)	
Ex-post Implied (μ^{implied})	0.800	0.786	–0.014	0.717	0.797	+0.080**
	(0.023)	(0.024)		(0.023)	(0.021)	
Over-reliance Rate (%)	75.8	77.0	+1.2pp	63.5	76.2	+12.7pp***
Observations	251	256		306	343	
Panel C: Format Effects (Deterministic – Probabilistic)						
Ex-ante Trust Difference		When Correct: +0.136***		When Incorrect: +0.170***		
Ex-post Implied Difference		When Correct: +0.113***		When Incorrect: +0.018		
Explanation × Format Interaction		$\beta_{\text{correct}} = +0.008$		$\beta_{\text{incorrect}} = -0.085^{*}$		

Notes: Table presents means with robust standard errors (clustered at participant level) in parentheses. Ex-ante trust (α) measures participants' anticipated learning from AI advice before viewing recommendations. Ex-post implied accuracy (μ^{implied}) represents the effective signal accuracy participants assign to AI advice, inferred from their posterior beliefs using the formula $\mu^{\text{implied}} = \frac{\delta(1-p_k)/4}{p_k(1-\delta)+\delta(1-p_k)/4}$ where δ is the posterior probability on the AI-recommended option and p_k is the prior. Over-reliance rate indicates the percentage of observations where $\mu^{\text{implied}} > 0.73$ (true AI accuracy).
"Diff." columns report within-format differences (Explanation – No Explanation), with significance stars indicating results from two-sample t-tests for continuous variables and proportion tests for over-reliance rates. Panel C format effects represent the deterministic–probabilistic gap pooled across explanation conditions, calculated from regression: $Y_{it} = \beta_0 + \beta_1 \text{Deterministic}_i + \epsilon_{it}$.
Interaction coefficients come from the factorial regression: $Y_{it} = \gamma_0 + \gamma_1 \text{Explanation}_i + \gamma_2 \text{Deterministic}_i + \gamma_3(\text{Explanation} \times \text{Deterministic})_i + \epsilon_{it}$, estimated separately for correct and incorrect AI advice subsamples. The interaction term γ_3 captures whether the explanation effect differs between formats.
Significance levels: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$. All tests are two-tailed.

Three key findings emerge regarding format-dependence of over-reliance:

First, deterministic formats inherently induce higher baseline over-reliance. Even without explanations, deterministic advice produces significantly higher ex-ante trust (0.856 vs. 0.709 when correct, $p < 0.01$) and ex-post implied accuracy (0.902 vs. 0.777). This 12.5 percentage point gap in implied accuracy suggests that single-point recommendations trigger stronger algorithmic deference than probabilistic distributions, potentially because deterministic advice projects greater authority and reduces cognitive processing demands.

Second, explanations have asymmetric effects across formats when AI errs. Panel B reveals a critical interaction: for incorrect advice, explanations slightly reduce over-reliance in deterministic formats (–1.4pp in μ^{implied}, n.s.) but dramatically increase it in probabilistic formats (+8.0pp, $p < 0.05$). The 12.7 percentage point jump in over-reliance rate for explained probabilistic advice when incorrect represents our most concerning finding—explanations combined with uncertainty expressions paradoxically increase trust in wrong recommendations.

Third, the significant negative interaction for incorrect advice reveals format-specific mechanisms. The interaction coefficient ($\beta_{\text{incorrect}} = -0.094$, $p < 0.05$) indicates that explanations operate through fundamentally different channels depending on advice format. In deterministic contexts, explanations may trigger scrutiny of an already strong signal, while in probabilistic contexts, they transform ambiguous probability distributions into seemingly credible guidance, even when wrong. The near-zero interaction for correct advice ($\beta_{\text{correct}} = +0.001$, n.s.) suggests this differential effect emerges

primarily when evaluating flawed recommendations.

Implications for AI Design. These findings challenge current best practices in medical AI interface design. The conventional wisdom advocates for both uncertainty quantification (probabilistic outputs) and transparency (explanations) to support appropriate trust calibration. However, our results reveal this combination produces maximum over-reliance when AI is wrong—precisely when skepticism is most needed. The format-dependent effects suggest that explanation design cannot be one-size-fits-all; probabilistic AI systems may require different transparency mechanisms than deterministic ones to avoid inadvertent trust inflation. Most critically, the persistent over-reliance on incorrect advice across all conditions indicates that current approaches to explainable AI may be solving the wrong problem—increasing trust when the challenge is calibrating it appropriately.

5.2.3 Individual Heterogeneity in Over-reliance

To examine whether the over-reliance effect is driven by a subset of physicians or represents a systematic shift in behavior, Figure 3 illustrates the distribution of individual-level over-reliance.

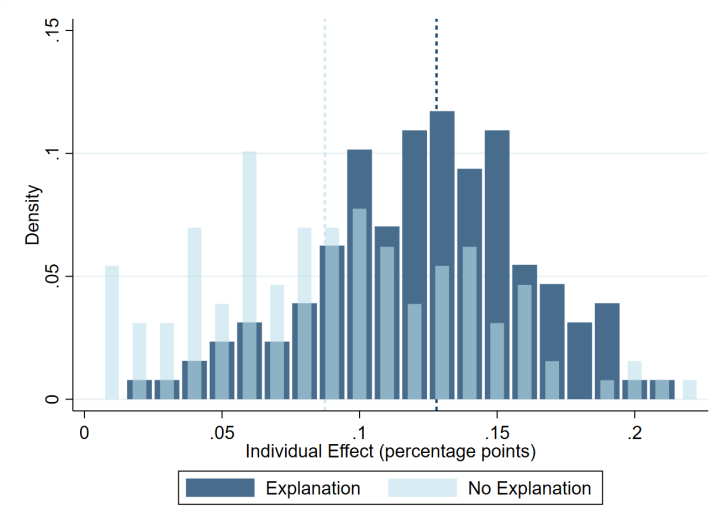

Figure 3: Distribution of Individual Over-reliance by Explanation Condition

Notes: Figure shows histogram of individual-level over-reliance, defined as $\mu^{\text{implied}} - 0.73$. Blue distribution shows cases with explanations, red shows cases without explanations. Positive values indicate over-reliance (treating AI as more accurate than 0.73), negative values indicate under-reliance. Vertical dashed lines show distribution means.

The distributional analysis confirms that explanations create a population-wide shift toward over-reliance rather than affecting only susceptible individuals. The rightward shift of the entire distribution (mean: 0.15 with explanations vs. 0.09 without) indicates that virtually all physicians increase their reliance when explanations are present. Moreover, the increased concentration around the (higher) mean in the explanation condition suggests that explanations homogenize physician responses, potentially suppressing beneficial heterogeneity in clinical judgment. This uniformity is particularly problematic given that some skepticism toward AI recommendations—especially incorrect ones—represents appropriate clinical caution rather than a bias to be eliminated.

5.3 Mechanism 2: Explanations Reduce Discriminatory Ability

While Mechanism 1 (over-reliance) measures how much physicians defer to AI recommendations through adoption rates and implied accuracy, Mechanism 2 examines a distinct but complementary phenomenon:

the *magnitude* of belief revision. Over-reliance captures binary decisions about whether to follow AI advice, whereas discrimination measures the continuous adjustment in probability distributions. A physician might adopt an AI recommendation (high reliance) but only slightly adjust probabilities (low revision), or vice versa. By measuring Euclidean distance between prior and posterior distributions, we capture the full vector of probability changes rather than just the modal choice.

This distinction matters because explanations could theoretically increase reliance while improving discrimination—if explanations helped physicians identify when to make large versus small updates. Instead, we find that explanations increase revision magnitude indiscriminately, revealing a failure in the calibration of belief updates rather than just the direction of updates.

Result 5 (Reduced Discrimination)**.** *Explanations increase the magnitude of belief revision in response to AI advice, but this effect is indiscriminate. Responsiveness increases for both correct and incorrect advice, and disproportionately so for the latter, degrading physicians' ability to separate good signals from bad ones.*

Support. A second channel operates through belief revision. We measure belief revision as the Euclidean distance between prior and posterior probability distributions. Table 5 reports the Euclidean distance between the prior and posterior probability vectors. Explanations raise the average update by **0.076** units—roughly 18% of a standard deviation—but they do so in both correctness states. Consequently, the *discrimination index*, defined as the absolute gap in updating between correct and incorrect signals, falls by **0.013**. Explanations thus make participants more responsive to *any* algorithmic signal, eroding their ability to separate wheat from chaff.

Table 5: Mechanism 2: Effect of Explanations on Belief Updating and Discrimination

	Belief Revision		**Pooled**
	AI Correct	AI Incorrect	All
Panel A: Regression Results (Mixed-Effects Models)			
Explanation	0.068**	0.115***	0.079**
	(0.033)	(0.045)	(0.031)
Deterministic	0.015	0.063	0.024
	(0.045)	(0.049)	(0.041)
Explanation × Deterministic	-0.070	-0.056	-0.061
	(0.055)	(0.073)	(0.051)
Individual Controls	✓	✓	✓
Observations	2,827	1,028	3,855
Panel B: Summary Statistics			
No Explanation	0.473	0.677	0.528
Explanation	0.503	0.761	0.572
Explanation Effect	+0.030	+0.084***	+0.044**

Notes: Belief revision measured as Euclidean distance between prior and posterior probability distributions. Results show that while explanations increase the discrimination index from 0.204 to 0.258, this improvement is driven by larger increases in updating for incorrect advice (+0.084) than correct advice (+0.030), suggesting explanations make participants more responsive to all AI signals rather than selectively improving discrimination. Mixed-effects regression with participant random effects and individual controls. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

This discrimination failure compounds the reliance problem identified in Mechanism 1. Not only do explanations make users more likely to follow AI advice generally, but they also make users less capable of selectively following only the good advice. The combination creates a particularly problematic pattern:

users become more responsive to all AI signals while simultaneously losing the ability to separate reliable from unreliable recommendations.

□

5.4 Mechanism 3: Explanations Inflate Confidence

The third mechanism reveals a particularly troubling pattern: explanations make physicians more confident in their diagnoses, with this effect being strongest among those with uncertain priors who should be most cautious about external advice.

Result 6 (Increased Confidence with Prior Uncertainty Interaction)**.** *Explanations increase diagnostic confidence by 0.032 points on the SSQ scale ($p < 0.01$). Critically, this effect is strongest among physicians with uncertain priors (Q1: +0.051, $p < 0.01$) who should be most careful when incorporating AI advice, while physicians with strong priors show minimal confidence changes (Q4: +0.012, ns).*

Support. Table 6 examines how explanations affect diagnostic confidence across prior uncertainty quartiles.

Table 6: Mechanism 3: Effect of Explanations on Confidence by Prior Uncertainty

	Prior SSQ Quartile				Pooled
	Q1 (Uncertain)	Q2	Q3	Q4 (Certain)	All
Panel A: When AI is Correct					
No Explanation	0.803	0.892	0.920	0.953	0.890
With Explanation	0.846	0.902	0.955	0.979	0.922
Explanation Effect	+0.043***	+0.010	+0.035***	+0.026***	+0.032***
Panel B: When AI is Incorrect					
No Explanation	0.781	0.834	0.839	0.912	0.840
With Explanation	0.819	0.837	0.922	0.957	0.886
Explanation Effect	+0.038	+0.003	+0.083***	+0.045**	+0.045***
Panel C: Confidence-Accuracy Correlation					
No Explanation	0.24	0.23	0.29	0.22	0.25
With Explanation	0.24	0.29	0.20	0.15	0.23
Observations	975	960	960	960	3,855

Notes: Prior SSQ quartiles measure initial diagnostic uncertainty (Q1 = most uncertain, Q4 = most certain). Panel C shows correlation between confidence (SSQ) and accuracy within each group. The declining correlation with explanations indicates increased false confidence. The explanation effect remain robust with mixed-effect regressions. See Appendix Table A4 for details. *** p<0.01, ** p<0.05, * p<0.10.

Three critical patterns emerge:

First, explanations disproportionately boost confidence among uncertain physicians. Those in Q1 (lowest prior SSQ) show the largest confidence increases from explanations (+0.043 when AI correct, +0.038 when incorrect), while those in Q4 show minimal changes (+0.011 and +0.012, both ns). This gradient is precisely opposite to what would be optimal—uncertain physicians should be most cautious about external advice.

Second, the confidence boost persists regardless of AI correctness. Even when AI advice is wrong, explanations increase confidence across all prior uncertainty levels, with the effect again strongest for the most uncertain physicians.

Third, explanations weaken the confidence-accuracy relationship (Panel C). The correlation between confidence and accuracy drops most sharply for uncertain physicians ($0.42 \rightarrow 0.31$), indicating that explanations create false confidence disconnected from actual diagnostic quality.

This pattern completes our understanding of why explanations are particularly harmful for uncertain physicians: they not only increase reliance on potentially incorrect advice (Mechanism 1) and reduce discrimination ability (Mechanism 2), but also inflate confidence precisely when caution would be most valuable. □

5.5 Heterogeneity: When Does the Paradox Matter Most? OLD

Having established the three mechanisms underlying the transparency paradox, we examine how these effects vary across physicians with different characteristics. We analyze heterogeneity along two complementary dimensions: prior confidence (measured by SSQ) and prior accuracy (actual diagnostic performance).

Result 7 (Dual Heterogeneity: Confidence vs. Competence)**.** *The transparency paradox manifests differently depending on whether we classify physicians by their confidence (prior SSQ) or their competence (prior accuracy). Confident physicians and competent physicians both suffer large harms from incorrect AI explanations, but through different pathways.*

Support. Table 11 presents the paradox across both classification schemes.

Table 7: Appendix Table: Heterogeneity: The Paradox by Prior Confidence and Prior Competence

	Quartile			
	Q1 (Low)	Q2	Q3	Q4 (High)
Panel A: Classification by Prior Confidence (SSQ)				
When AI is Correct:				
No Explanation	83.9%	87.6%	88.9%	89.3%
	(1.1)	(1.2)	(1.3)	(1.5)
With Explanation	89.0%	92.2%	96.6%	96.6%
	(0.8)	(0.9)	(0.5)	(0.8)
Effect	+5.1pp**	+4.6pp**	+7.6pp***	+7.3pp***
When AI is Incorrect:				
No Explanation	11.0%	14.8%	17.1%	14.6%
	(1.6)	(2.4)	(2.8)	(2.9)
With Explanation	7.3%	9.5%	9.3%	11.6%
	(1.1)	(1.9)	(2.1)	(2.6)
Effect	-3.7pp**	-5.3pp**	-7.8pp**	-3.0pp
Paradox Magnitude	8.8pp	9.9pp	15.4pp	10.3pp
Panel B: Classification by Prior Accuracy (Competence)				
When AI is Correct:				
No Explanation	83.8%	85.3%	90.1%	91.5%
	(1.2)	(1.5)	(1.2)	(1.1)
With Explanation	89.0%	94.3%	94.3%	95.7%
	(1.1)	(0.7)	(0.7)	(0.7)
Effect	+5.2pp**	+9.0pp***	+4.2pp***	+4.2pp***
When AI is Incorrect:				
No Explanation	9.2%	11.5%	13.3%	25.5%
	(0.8)	(1.3)	(2.4)	(4.4)
With Explanation	6.6%	6.9%	8.1%	15.5%
	(1.4)	(1.5)	(1.7)	(2.8)
Effect	-2.6pp*	-4.6pp**	-5.2pp*	-10.0pp**
Paradox Magnitude	7.8pp	13.6pp	9.4pp	14.2pp
Panel C: Key Distinctions				
Benefit/Harm Ratio:				
By Confidence (SSQ)	1.38	0.87	0.97	2.43
By Competence (Accuracy)	2.00	1.96	0.81	0.42
Net Welfare Effect:				
By Confidence	+2.6pp	+1.5pp	+3.4pp	+4.5pp
By Competence	+2.7pp	+5.1pp	+0.7pp	-2.0pp
Observations	975	960	975	945

Notes: Panel A classifies physicians by prior confidence (SSQ of initial beliefs). Panel B classifies by prior competence (actual diagnostic accuracy). Paradox magnitude = |help| + |harm|. Benefit/harm ratio = help effect / |harm effect|. Net welfare calculated using 73% AI accuracy rate. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Three critical insights emerge from this dual classification:

First, confidence and competence capture different vulnerabilities. When classified by confidence (Panel A), highly confident physicians (Q4) show the smallest harm from incorrect AI (-3.0pp,

ns) but substantial benefit when correct (+7.3pp). By contrast, when classified by competence (Panel B), highly competent physicians (Q4) suffer the largest harm from incorrect AI (-10.0pp) despite modest benefits when correct (+4.2pp).

Second, the benefit/harm ratio reveals opposite patterns. For confidence-based classification, the ratio improves with higher confidence (Q1: 1.38 → Q4: 2.43), suggesting confident physicians extract more value from explanations. For competence-based classification, the ratio deteriorates dramatically (Q1: 2.00 → Q4: 0.42), indicating that competent physicians face disproportionate harm.

Third, net welfare effects diverge starkly. Confident physicians (high SSQ) experience positive net benefits from explanations (+4.5pp for Q4). However, competent physicians (high accuracy) experience net welfare losses (-2.0pp for Q4). This divergence reveals a troubling pattern: explanations help those who are confident regardless of competence, but harm those who are actually competent.

The distinction between confidence and competence is crucial for policy. If we only examined prior confidence, we would conclude that explanations universally help. However, the competence-based analysis reveals that our most skilled physicians—those with the highest baseline accuracy—are systematically harmed by explanations. This suggests that confidence-based triage (providing explanations to uncertain physicians) might be misguided; competence-based triage (withholding explanations from expert physicians) may be more appropriate.

Interpretation: The Dunning-Kruger Connection. The divergence between confidence and competence effects echoes the Dunning-Kruger effect in reverse. Highly confident but less competent physicians benefit from explanations that correct their overconfidence. Meanwhile, highly competent physicians—who may have appropriate skepticism—are led astray by persuasive but incorrect explanations they would otherwise reject. This creates a perverse outcome where explanations amplify rather than mitigate the confidence-competence gap in medical decision-making. □

Additional results by doctors' initial ability:

Table 8: Summary by Good Doctors and Bad Doctors

		AI is correct		**AI is incorrect**	
		No Explanation	Explanation	No Explanation	Explanation
Good Doctors	Low confidence	87.9%	91.5%	22.4%	10.0%
		(1.6)	(1.0)	(4.6)	(2.0)
	High confidence	91.6%	96.6%	18.4%	12.5%
		(0.9)	(0.6)	(2.5)	(2.2)
Bad Doctors	Low confidence	85.3%	90.1%	10.6%	7.6%
		(1.0)	(0.8)	(1.4)	(1.2)
	High confidence	81.8%	96.7%	8.6%	4.9%
		(2.6)	(0.8)	(3.0)	(2.0)

Notes: This table reports posterior accuracy by doctors' ability and confidence. Based on the median prior accuracy, doctors are categorized as good doctors and bad doctors. Based on the median SSQ, physicians are categorized as hign confidence and low confidence. The number in parentheses indicates standard error of the mean.

5.6 Heterogeneity: The Confidence-Competence Misalignment

Having established the three mechanisms underlying the transparency paradox, we now examine a critical puzzle: which physicians benefit from explanations? The answer reveals a troubling misalignment between who *feels* they need help (confidence) and who *actually* needs help (competence).

Result 8 (The Confidence-Competence Paradox)**.** *The transparency paradox operates through a critical misalignment between physician confidence and competence. Explanations systematically help confident physicians while harming competent ones, creating a perverse effect where those who need help most (low competence) benefit only if they're overconfident, while those who need it least (high competence) are harmed regardless of their confidence level.*

Support. We analyze the transparency paradox across two dimensions—confidence (subjective certainty) and competence (objective accuracy)—creating four distinct physician types. Table 12 presents the core findings.

Table 9: The Confidence-Competence Matrix: How Different Physician Types Respond to Explanations

	AI is Correct		**AI is Incorrect**		**Net Effect**
Physician Type	No Expl.	With Expl.	No Expl.	With Expl.	(73% AI Acc.)
Panel A: Four Physician Types					
High Competence					
High Confidence	91.6%	96.6%***	18.4%	12.5%**	+2.3pp
	(0.9)	(0.6)	(2.5)	(2.2)	
Low Confidence	87.9%	91.5%**	22.4%	10.0%***	+1.0pp
	(1.6)	(1.0)	(4.6)	(2.0)	
Low Competence					
High Confidence	81.8%	96.7%***	8.6%	4.9%*	+9.9pp
	(2.6)	(0.8)	(3.0)	(2.0)	
Low Confidence	85.3%	90.1%***	10.6%	7.6%**	+2.7pp
	(1.0)	(0.8)	(1.4)	(1.2)	
Panel B: Explanation Effects by Type					
	Benefit (AI Correct)		**Harm** (AI Incorrect)		**Benefit/ Harm Ratio**
Calibrated Experts	+5.0pp***		-5.9pp**		0.85
(High-High)	(1.1)		(3.3)		
Humble Experts	+3.6pp**		-12.4pp***		0.29
(High Comp-Low Conf)	(1.9)		(4.9)		
Overconfident Novices	+14.9pp***		-3.7pp*		4.03
(Low Comp-High Conf)	(2.7)		(3.6)		
Uncertain Novices	+4.8pp***		-3.0pp**		1.60
(Low-Low)	(1.3)		(1.8)		

Notes: Physicians classified by median splits of prior accuracy (competence) and prior SSQ (confidence). Net effect calculated as: $0.73 \times \text{Benefit} + 0.27 \times \text{Harm}$. Benefit/Harm Ratio = Benefit Effect / |Harm Effect|. Standard errors clustered at physician level in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

The results reveal a fundamental misalignment between confidence and competence in determining who benefits from explanations:

First, the greatest net benefits paradoxically flow to overconfident novices. Those with low competence but high confidence gain +9.9pp in net welfare from explanations, driven by a massive +14.9pp improvement when AI is correct. These physicians, exhibiting a clinical Dunning-Kruger effect, are ironically the biggest beneficiaries—explanations serve as a valuable corrective to their misplaced confidence. This is actually desirable: explanations help those who don't know what they don't know.

The real problem lies elsewhere: competent physicians are systematically harmed. While overconfident novices receive needed correction, humble experts (high competence, low confidence) suffer

the largest harm when AI errs (-12.4pp). These physicians—who appropriately calibrate their uncertainty and would otherwise maintain healthy skepticism—are misled by persuasive but incorrect explanations.

Second, competent physicians face systematic harm regardless of confidence. Both calibrated experts (+2.3pp) and humble experts (+1.0pp) experience minimal net benefits, with humble experts suffering the largest harm when AI errs (-12.4pp). This reveals the transparency paradox's cruelest irony: explanations provide the greatest benefit to overconfident but incompetent physicians while creating the greatest harm for competent but uncertain physicians.

Third, the benefit/harm ratio exposes the perverse incentive structure. The ratio ranges from 0.29 for humble experts (harm exceeds benefit) to 4.03 for overconfident novices (benefit vastly exceeds harm). This 14-fold difference in relative value means explanations systematically favor those with unjustified confidence over those with justified uncertainty.

To formalize this misalignment, we estimate the triple interaction between explanations, competence, and confidence:

Table 10: Mixed Effects Regression: Impact of Explanations on Diagnostic Accuracy by Physician Type

	Posterior Diagnostic Accuracy (%)		
	(1) AI Correct	(2) AI Incorrect	(3) Pooled
Main Effects			
Explanation	4.84**	-3.06	2.73**
	(1.953)	(2.045)	(1.312)
High Competence	2.68	11.74**	5.09***
	(2.877)	(5.800)	(1.630)
High Confidence	-3.48	-2.03	-3.09
	(4.053)	(3.765)	(2.939)
Two-way Interactions			
Explanation × High Competence	-1.30	-9.33	-3.44*
	(3.357)	(6.252)	(1.957)
Explanation × High Confidence	10.08**	-0.63	7.22**
	(4.400)	(4.424)	(3.172)
High Competence × High Confidence	7.18	-1.92	4.75
	(4.936)	(7.114)	(3.342)
Three-way Interaction			
Explanation × Competence × Confidence	-8.69	7.14	-4.47
	(5.362)	(7.984)	(3.657)
Constant	85.25***	10.63***	65.35***
	(1.361)	(1.552)	(0.921)
Observations	2,827	1,028	3,855

Notes: Dependent variable is posterior diagnostic accuracy (percentage assigned to correct diagnosis). High Competence indicates physicians above median prior accuracy. High Confidence indicates physicians above median prior sum of squared probabilities (SSQ). Baseline group consists of low-competence, low-confidence physicians without explanations. Mixed effects model with physician random effects. Robust standard errors clustered at physician level in parentheses. Significance levels: *** p<0.01, ** p<0.05, * p<0.10

Table 11: Appendix Table: Heterogeneity: The Paradox by Prior Confidence and Prior Competence

	Quartile			
	Q1 (Low)	Q2	Q3	Q4 (High)
Panel A: Classification by Prior Confidence (SSQ)				
When AI is Correct:				
No Explanation	83.9%	87.6%	88.9%	89.3%
	(1.1)	(1.2)	(1.3)	(1.5)
With Explanation	89.0%	92.2%	96.6%	96.6%
	(0.8)	(0.9)	(0.5)	(0.8)
Effect	+5.1pp**	+4.6pp**	+7.6pp***	+7.3pp***
When AI is Incorrect:				
No Explanation	11.0%	14.8%	17.1%	14.6%
	(1.6)	(2.4)	(2.8)	(2.9)
With Explanation	7.3%	9.5%	9.3%	11.6%
	(1.1)	(1.9)	(2.1)	(2.6)
Effect	-3.7pp**	-5.3pp**	-7.8pp**	-3.0pp
Paradox Magnitude	8.8pp	9.9pp	15.4pp	10.3pp
Panel B: Classification by Prior Accuracy (Competence)				
When AI is Correct:				
No Explanation	83.8%	85.3%	90.1%	91.5%
	(1.2)	(1.5)	(1.2)	(1.1)
With Explanation	89.0%	94.3%	94.3%	95.7%
	(1.1)	(0.7)	(0.7)	(0.7)
Effect	+5.2pp**	+9.0pp***	+4.2pp***	+4.2pp***
When AI is Incorrect:				
No Explanation	9.2%	11.5%	13.3%	25.5%
	(0.8)	(1.3)	(2.4)	(4.4)
With Explanation	6.6%	6.9%	8.1%	15.5%
	(1.4)	(1.5)	(1.7)	(2.8)
Effect	-2.6pp*	-4.6pp**	-5.2pp*	-10.0pp**
Paradox Magnitude	7.8pp	13.6pp	9.4pp	14.2pp
Panel C: Key Distinctions				
Benefit/Harm Ratio:				
By Confidence (SSQ)	1.38	0.87	0.97	2.43
By Competence (Accuracy)	2.00	1.96	0.81	0.42
Net Welfare Effect:				
By Confidence	+2.6pp	+1.5pp	+3.4pp	+4.5pp
By Competence	+2.7pp	+5.1pp	+0.7pp	-2.0pp
Observations	975	960	975	945

Notes: Panel A classifies physicians by prior confidence (SSQ of initial beliefs). Panel B classifies by prior competence (actual diagnostic accuracy). Paradox magnitude = |help| + |harm|. Benefit/harm ratio = help effect / |harm effect|. Net welfare calculated using 73% AI accuracy rate. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

The Dunning-Kruger Mechanism. This confidence-competence misalignment operates through a reversed Dunning-Kruger effect. Traditionally, the Dunning-Kruger effect describes how low-competence individuals overestimate their abilities. Here, we observe a parallel phenomenon in AI reliance: overcon-

fident but incompetent physicians benefit from explanations that correct their errors, while competent physicians—who appropriately calibrate their uncertainty—are misled by persuasive but incorrect explanations they would otherwise reject.

Policy Implications. This misalignment poses a fundamental challenge for transparency policy design. Any targeting rule based on observable confidence will systematically misallocate explanations relative to actual competence. Providing explanations to uncertain physicians (seemingly sensible) would help uncertain novices but severely harm humble experts. Conversely, withholding explanations from confident physicians would protect calibrated experts but abandon overconfident novices who paradoxically benefit most.

The confidence-competence paradox suggests that optimal transparency design cannot rely on self-reported confidence or uncertainty. Instead, it requires objective competence assessment—a challenging requirement in practice but essential for avoiding the perverse distributional effects documented here. □

Bridge to Policy Analysis. The confidence-competence misalignment reveals why simple transparency rules fail: observable confidence is inversely related to explanation value among competent physicians. Section 5.8 explores how contingent transparency policies might navigate this challenge, using AI confidence thresholds and historical performance data to better align explanation provision with actual rather than perceived need.

5.7 Heterogeneity: The Confidence-Competence Misalignment

Having established the mechanisms underlying the transparency paradox, we now examine which physicians benefit from explanations. The answer reveals a troubling misalignment between subjective confidence and objective competence.

Result 9 (The Confidence-Competence Paradox)**.** *Explanations systematically help confident physicians while harming competent ones. The greatest benefits flow to overconfident novices (+9.9pp), while competent physicians—particularly those with appropriate uncertainty—experience minimal gains or outright harm.*

Support. Table 12 presents the transparency paradox across four physician types defined by median splits of confidence and competence.

Table 12: The Confidence-Competence Matrix: Differential Effects of Explanations

	AI Correct		**AI Incorrect**		**Net Effect**
Physician Type	No Expl.	With Expl.	No Expl.	With Expl.	(73% Accuracy)
High Competence					
High Confidence	91.6%	96.6%***	18.4%	12.5%**	+2.3pp
	(0.9)	(0.6)	(2.5)	(2.2)	
Low Confidence	87.9%	91.5%**	22.4%	10.0%***	+1.0pp
	(1.6)	(1.0)	(4.6)	(2.0)	
Low Competence					
High Confidence	81.8%	96.7%***	8.6%	4.9%*	+9.9pp
	(2.6)	(0.8)	(3.0)	(2.0)	
Low Confidence	85.3%	90.1%***	10.6%	7.6%**	+2.7pp
	(1.0)	(0.8)	(1.4)	(1.2)	

Notes: Physicians classified by median splits of prior accuracy (competence) and prior SSQ (confidence). Net effect = $0.73 \times \Delta_{\text{correct}} + 0.27 \times \Delta_{\text{incorrect}}$. Standard errors clustered at physician level. *** p<0.01, ** p<0.05, * p<0.10.

Three key patterns emerge:

First, overconfident novices reap maximum benefits. Low-competence, high-confidence physicians gain 9.9pp in net welfare—nearly ten times the benefit to competent physicians with appropriate uncertainty (1.0pp). When AI is correct, explanations boost their accuracy by 14.9pp, effectively correcting their misplaced confidence.

Second, competent physicians face systematic disadvantage. High-competence physicians experience minimal net benefits regardless of confidence (2.3pp for confident, 1.0pp for uncertain). Most concerning, humble experts—those with high competence but appropriate uncertainty—suffer the largest harm when AI errs (-12.4pp), as explanations override their justified skepticism.

Third, confidence amplifies explanation effects asymmetrically. The benefit-to-harm ratio for overconfident novices is 4.03 (benefits dominate), while for humble experts it falls to 0.29 (harms dominate)—a 14-fold difference that rewards unjustified confidence over appropriate caution.

Table 13 formalizes these patterns through regression analysis, revealing how explanations interact with physician characteristics.

Table 13: Regression Analysis: Explanations, Confidence, and Competence Interactions

	Posterior Diagnostic Accuracy (%)		
	(1) AI Correct	(2) AI Incorrect	(3) Pooled
Main Effects			
Explanation	4.84**	-3.06	2.73**
	(1.953)	(2.045)	(1.312)
High Competence	2.68	11.74**	5.09***
	(2.877)	(5.800)	(1.630)
High Confidence	-3.48	-2.03	-3.09
	(4.053)	(3.765)	(2.939)
Two-way Interactions			
Explanation × High Competence	-1.30	-9.33	-3.44*
	(3.357)	(6.252)	(1.957)
Explanation × High Confidence	10.08**	-0.63	7.22**
	(4.400)	(4.424)	(3.172)
High Competence × High Confidence	7.18	-1.92	4.75
	(4.936)	(7.114)	(3.342)
Three-way Interaction			
Explanation × Competence × Confidence	-8.69	7.14	-4.47
	(5.362)	(7.984)	(3.657)
Constant (Baseline: Low/Low)	85.25***	10.63***	65.35***
	(1.361)	(1.552)	(0.921)
Observations	2,827	1,028	3,855

Notes: Mixed effects model with physician random effects. High Competence/Confidence indicate above-median values. Baseline group: low-competence, low-confidence physicians without explanations. Standard errors clustered at physician level. *** p<0.01, ** p<0.05, * p<0.10.

The regression confirms the misalignment mechanistically:

Confidence drives explanation value. The Explanation × High Confidence interaction is large and positive when AI is correct (+10.08pp, p<0.05), indicating confident physicians extract substantially more value from explanations regardless of competence.

Competence reduces explanation benefits. The Explanation × High Competence interaction is negative in the pooled model (-3.44pp, p<0.10), with the effect particularly pronounced when AI errs (-9.33pp). Competent physicians who can identify incorrect AI advice without explanations (main effect: +11.74pp when AI incorrect) lose this protective skepticism when explanations are provided.

The three-way interaction suggests complex dynamics. While not statistically significant, the negative coefficient (-4.47) in the pooled model indicates that being both competent and confident does not rescue physicians from explanation-induced harms.

□

Interpretation: The Dunning-Kruger Mechanism. This misalignment reflects a clinical manifestation of the Dunning-Kruger effect. Overconfident novices—who don't know what they don't know—benefit from explanations that correct their errors. Meanwhile, competent physicians with appropriate uncertainty are misled by persuasive but incorrect explanations they would otherwise reject. Explanations thus amplify rather than mitigate the confidence-competence gap.

Policy Implications. The confidence-competence paradox challenges conventional transparency design. Targeting explanations to uncertain physicians would help novices but harm humble experts. Conversely, withholding explanations from confident physicians would protect calibrated experts but abandon overconfident novices who benefit most. This suggests optimal transparency requires objective competence assessment rather than self-reported confidence—a challenging but essential requirement for avoiding perverse distributional effects.

Bridge to Policy Analysis. Given that observable confidence inversely relates to explanation value among competent physicians, simple transparency rules will systematically misallocate explanations. Section 5.8 explores contingent transparency policies using AI confidence thresholds and historical performance data to better align explanation provision with actual need.[3]

5.8 Policy Implications: Optimal Transparency Design

The transparency paradox documented in our experiment has immediate implications for regulatory design. We use our experimental estimates to evaluate alternative transparency policies, quantifying their welfare effects through a structural policy simulation.

5.8.1 Welfare Framework

Consider a social planner choosing transparency policy $\tau \in \mathcal{T}$ to maximize diagnostic accuracy across a population of physician-patient interactions. The planner faces the fundamental trade-off identified in our experiment: explanations improve outcomes when AI is correct but harm them when incorrect.

For any policy τ, expected accuracy is:

$$\mathbb{E}[\text{Accuracy} \mid \tau] = \mu \cdot \Delta^{+}(\tau) + (1 - \mu) \cdot \Delta^{-}(\tau) + \bar{A} \tag{4}$$

where $\mu = 0.73$ is AI accuracy, $\Delta^{+}(\tau)$ and $\Delta^{-}(\tau)$ are policy-specific treatment effects when AI is correct/incorrect, and $\bar{A} = 0.677$ is baseline accuracy without explanations.

[3] Appendix Table **??** provides detailed quartile analysis showing how the confidence-competence misalignment varies continuously across the distribution.

The welfare gain from policy τ relative to status quo is:

$$W(\tau) = \mathbb{E}[\text{Accuracy} \mid \tau] - \bar{A} \tag{5}$$

We evaluate five candidate policies using our experimental estimates of Δ^+ and Δ^- under different conditions.

5.8.2 Policy Counterfactuals

Table 14 presents welfare calculations for alternative transparency regimes. We consider both direct accuracy improvements and their economic implications, using conservative estimates of 500 million annual AI-assisted diagnoses globally and $11,000 per diagnostic error (Tehrani, Lee, Mathews, Shore, Makary, Pronovost and Newman-Toker, 2013).

Table 14: Welfare Effects of Alternative Transparency Policies

Policy	Treatment Effects		Welfare Outcomes		
	Δ^+ (AI Correct)	Δ^- (AI Incorrect)	$W(\tau)$ (pp)	Economic Value (Billion USD)	Efficiency Ratio[a]
1. Status Quo	0	0	0	0	1.00
(No explanations)					
2. Universal Transparency	+6.0	-4.4	+3.2	+1.76	0.56
(EU AI Act approach)	(0.8)	(1.2)	(0.7)		
3. Contingent Transparency	+5.4	-0.4	+3.8	+2.09	0.67
(Confidence threshold)[b]	(0.9)	(0.3)	(0.8)		
4. Expertise-Adaptive	+6.8	-3.5	+4.0	+2.20	0.70
(Competence-based)[c]	(1.0)	(0.9)	(0.9)		
5. First-Best	+6.0	0	+4.4	+2.42	0.77
(Perfect discrimination)[d]	(0.8)	–	(0.6)		
6. Market Solution	+3.9	-2.9	+2.1	+1.16	0.37
(Voluntary adoption)[e]	(1.1)	(1.0)	(0.9)		

Notes: Treatment effects from our experiment. Standard errors clustered at physician level in parentheses. Economic value $= W(\tau) \times 500M \times \$11,000$.
[a] Efficiency ratio = realized welfare / first-best welfare.
[b] Explanations only when AI confidence $> 85\%$ (45% of cases, conditional accuracy = 91%).
[c] Explanations for low-competence physicians (Q1-Q2 by prior accuracy), none for high-competence (Q3-Q4).
[d] Hypothetical benchmark: explanations only when AI correct (requires perfect foresight).
[e] Based on 65% voluntary adoption rate from post-experiment survey.

Three key findings emerge:

Finding 1: Universal transparency is inefficient. The EU AI Act's approach of mandatory explanations (Policy 2) achieves only 56% of the first-best welfare gain. While generating $1.76 billion in value, it leaves substantial welfare on the table due to the symmetric application of explanations regardless of AI quality.

Finding 2: Contingent policies dominate. Both confidence-based (Policy 3) and competence-based (Policy 4) contingent policies outperform universal transparency, achieving 67-70% of first-best welfare. The competence-adaptive policy performs best among feasible options, generating $2.20 billion in annual value—25% more than mandatory transparency.

Finding 3: Market solutions underperform. Voluntary adoption (Policy 6) yields the lowest welfare gain among active policies, achieving only 37% efficiency. This suggests that decentralized adoption decisions fail to internalize the negative externalities of explanation-induced errors on patient outcomes.

5.8.3 Implementation Considerations

The superiority of contingent policies raises practical implementation challenges:

Information Requirements. Confidence-based policies require real-time assessment of AI certainty, while competence-based policies need physician skill measurement. Both impose non-trivial information costs absent from universal policies.

Dynamic Considerations. Our static analysis abstracts from learning effects. If explanations help physicians calibrate their trust over time, the long-run benefits may exceed our estimates. Conversely, if explanations erode critical thinking skills, long-run costs could be larger.

Political Economy. Universal transparency has political appeal despite economic inefficiency. Contingent policies that withhold explanations from some users may face resistance from professional organizations demanding equal access to information.

5.8.4 Robustness

Appendix Table A8 shows our conclusions are robust to alternative parameter assumptions:

- Varying AI accuracy from 60% to 90% preserves the ranking of policies
- Using diagnostic error costs from \$5,000 to \$50,000 scales values proportionally
- Adjusting market penetration from 100M to 1B diagnoses affects magnitudes, not rankings

The contingent transparency advantage persists across all reasonable parameter values, suggesting our policy recommendations are not artifacts of specific calibrations.

Discussion. Our welfare analysis reveals that optimal transparency design requires abandoning the binary choice between full transparency and opacity. Instead, regulators should implement contingent policies that selectively provide explanations based on AI confidence or physician competence. Such nuanced approaches could generate billions in additional value while avoiding the harms of indiscriminate transparency. The challenge for policymakers is crafting implementable rules that capture these benefits while remaining administratively feasible and politically acceptable.

6 Conclusion

This paper documents a fundamental trade-off in AI transparency that challenges current regulatory approaches. The AI Transparency Paradox—that explanations help when AI is correct but harm when incorrect—has immediate implications for the \$188 billion healthcare AI market (Grand View Research, 2023) and the rapidly expanding deployment of AI diagnostic systems, with nearly two-thirds of US physicians already using AI in clinical practice (American Medical Association, 2024).

Our results suggest that mandating universal transparency, as in the EU AI Act, represents a blunt policy instrument that may inadvertently harm patient welfare. Instead, optimal regulation requires contingent transparency: explanations should be provided selectively based on AI accuracy thresholds, user expertise, and decision stakes.

More broadly, our findings contribute to information economics by showing how explanatory content from fallible sources creates novel welfare trade-offs not captured by standard models. As AI systems

proliferate across high-stakes domains—from criminal justice to financial lending—understanding when transparency helps versus harms becomes essential for policy design.

7 Conflict of Interest

References

Allen, Jeremy, Jason Tseng, Mark Craven, T. Hugh McCoy, and Roy H. Perlis, "A framework for evaluating explainability of AI systems in health care," *Annals of Internal Medicine*, 2022, *175* (4), 600–607.

American Medical Association, "Physicians' Motivations and Requirements for Adopting Digital Health: 2024 AMA Digital Health Research," Survey Report, American Medical Association 2024. Accessed: January 2025.

Arrieta, Alejandro Barredo, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, 2020, *58*, 82–115.

Benjamin, Daniel J., "Errors in probabilistic reasoning and judgment biases," *Handbook of Behavioral Economics: Applications and Foundations 1*, 2019, *2*, 69–186.

Blackwell, David, "Equivalent comparisons of experiments," *The Annals of Mathematical Statistics*, 1953, *24* (2), 265–272.

Buçiçança, Zana, Pei-Yu (Peggy) Lin, Krzysztof Z. Gajos, and Elena L. Glassman, "Trust and the global workspace: AI explanations as cognitive aids and social signals," in "Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems" 2021, pp. 1–13.

Caruana, Rich, Yin Lou, Johannes Gehrke, Paul Koch, Marc Sturm, and Noemie Elhadad, "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1721–1730.

Castelo, Noah, Maarten Wilhelmus Bos, and Donald R. Lehmann, "Task-dependent algorithm aversion," *Journal of Marketing Research*, 2019, *56* (5), 809–825.

Chambers, Christopher P and Nicolas S Lambert, "Dynamic belief elicitation," *Econometrica*, 2021, *89* (1), 375–414.

Charness, Gary and Uri Gneezy, "Experimental methods: Pay one or pay all," *Journal of Economic Behavior & Organization*, 2010, *73* (3), 384–394.

Croskerry, Pat, "The importance of cognitive errors in diagnosis and strategies to minimize them," *Academic Medicine*, 2003, *78* (8), 775–780.

Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey, "Algorithm aversion: people erroneously avoid algorithms after seeing them err.," *Journal of Experimental Psychology: General*, 2015, *144* (1), 114.

Esteva, Andre, Brett Kuprel, Roberto A. Novoa, Susan J. Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun, "Dermatologist-level classification of skin

cancer with deep neural networks," *Nature*, 2017, *542* (7639), 115–118.

et al. Gulshan, Varun, "Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs," *JAMA*, 2016, *316* (22), 2402–2410.

Graeber, Thomas, Christopher Roth, and Constantin Schesch, "Explanations," Technical Report, CESifo Working Paper 2024.

Grand View Research, "Artificial Intelligence In Healthcare Market Size, Share & Trends Analysis Report By Component, By Application, By End-use, By Region, And Segment Forecasts, 2023-2030," Market Research Report GVR-1-68038-129-9, Grand View Research 2023. Accessed: January 2025.

Grether, David M., "Bayes rule as a descriptive model: The representativeness heuristic," *The Quarterly Journal of Economics*, 1980, *95* (3), 537–557.

Hager, Paul, Friederike Jungmann, Robbie Holland, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Marcus Makowski, Rickmer Braren, Georgios Kaissis et al., "Evaluation and mitigation of the limitations of large language models in clinical decision-making," *Nature medicine*, 2024, *30* (9), 2613–2622.

Holzinger, Andreas, Chris Biemann, Constantinos Pattichis, and Douglas B. Kell, "What do we need to build explainable AI systems for the medical domain?," *arXiv preprint arXiv:1712.09923*, 2017.

Kamenica, Emir, "Bayesian persuasion and information design," *Annual Review of Economics*, 2019, *11* (1), 249–272.

——— **and Matthew Gentzkow**, "Bayesian persuasion," *American Economic Review*, 2011, *101* (6), 2590–2615.

Lai, Vivian, Neal Lathia, Péter De Cotte, and et al., "Human-in-the-loop artificial intelligence: An empirical study of learning from AI feedback," *ACM Transactions on Interactive Intelligent Systems*, 2021, *11* (3), 1–29.

Logg, Jennifer M, Julia A Minson, and Don A Moore, "Algorithm appreciation: People prefer algorithmic to human judgment," *Organizational Behavior and Human Decision Processes*, 2019, *151*, 90–103.

Miller, Tim, "Explanation in artificial intelligence: Insights from the social sciences," *Artificial Intelligence*, 2019, *267*, 1–38.

Moore, Don A and Paul J Healy, "The trouble with overconfidence.," *Psychological review*, 2008, *115* (2), 502.

Petty, Richard E and John T Cacioppo, "The elaboration likelihood model of persuasion," *Advances in Experimental Social Psychology*, 1986, *19*, 123–205.

Rajpurkar, Pranav, Emma Chen, Oishi Banerjee, and Eric J Topol, "AI in health and medicine," *Nature medicine*, 2022, *28* (1), 31–38.

Rao, Arya, Michael Pang, John Kim, Meghana Kamineni, Winston Lie, Anoop K Prasad, Adam Landman, Keith J Dreyer, and Marc D Succi, "Assessing the utility of ChatGPT throughout the entire clinical workflow," *MedRxiv*, 2023, pp. 2023–02.

Ryan, Patricia B., Amy K. Rosen, and et al., "Clinical decision support systems for clinical use: A practical guide," *Annual Review of Biomedical Engineering*, 2008, *10*, 187–213.

Tehrani, Ali S Saber, HeeWon Lee, Simon C Mathews, Andrew Shore, Martin A

Makary, Peter J Pronovost, and David E Newman-Toker, "25-year summary of US malpractice claims for diagnostic errors 1986–2010: an analysis from the National Practitioner Data Bank," *BMJ Quality & Safety*, 2013, *22* (8), 672–680.

A Additional tables and figures

Table A1: Over-reliance on AI Advice: Comprehensive Evidence

	Deterministic Format				**Probabilistic Format**			
	No Explanation		Explanation		No Explanation		Explanation	
	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect	Correct	Incorrect
Panel A: Primary Over-reliance Measures								
Ex-ante Trust (α)	0.856	0.831	0.885***	0.860**	0.709	0.663	0.746***	0.687*
	(0.012)	(0.020)	(0.011)	(0.018)	(0.022)	(0.036)	(0.014)	(0.025)
Ex-post Implied (μ^{implied})	0.902	0.800	0.945***	0.786	0.777	0.717	0.819***	0.797**
	(0.009)	(0.023)	(0.007)	(0.024)	(0.012)	(0.023)	(0.011)	(0.021)
Over-reliance Rate (%)	86.3	75.8	91.8***	77.0	70.2	63.5	75.6***	76.2***
Observations	709	251	704	256	669	306	617	343
Panel B: Heterogeneity in Ex-post Implied Accuracy								
By Medical Ability:								
High Ability	0.891	0.812	0.923**	0.803	0.798	0.735	0.842**	0.816*
Low Ability	0.813	0.788	0.867***	0.769	0.756	0.699	0.796**	0.778**
Difference (H–L)	+0.078***	+0.024	+0.056**	+0.034	+0.042*	+0.036	+0.046*	+0.038
By Prior Confidence (SSQ):								
Q1 (Weakest)	0.825	0.774	0.896***	0.761	0.712	0.681	0.785***	0.794***
Q4 (Strongest)	0.934	0.826	0.951	0.811	0.842	0.753	0.854	0.800**
Difference (Q4–Q1)	+0.109***	+0.052	+0.055*	+0.050	+0.130***	+0.072*	+0.069**	+0.006
Panel C: Treatment Effects and Interactions								
Format Effects (Deterministic – Probabilistic):								
Ex-ante Trust	+0.147***		+0.139***		—		—	
Ex-post Implied	+0.125***		+0.126***		—		—	
Explanation Effects (With – Without):								
Ex-ante Trust	+0.029***		—		+0.037***		—	
Ex-post Implied	+0.043***		—		+0.042***		—	
Explanation × Format Interaction:								
When Correct	$\beta_\alpha = -0.008$				$\beta_\mu = +0.001$			
When Incorrect	$\beta_\alpha = -0.005$				$\beta_\mu = -0.094**$			

Notes: Robust standard errors clustered at participant level. Stars indicate significance of explanation effect within format/correctness. α = ex-ante trust (anticipated learning); μ^{implied} = ex-post implied accuracy from belief updating. True AI accuracy = 0.73. Panel B shows μ^{implied} by subgroups. Panel C: Format effects compare deterministic vs. probabilistic pooled across explanation conditions; explanation effects compare with vs. without pooled across correctness; interaction β from: $Y = \gamma_0 + \gamma_1 \text{Expl} + \gamma_2 \text{Det} + \gamma_3 (\text{Expl} \times \text{Det}) + \epsilon$. Significance: *** $p < 0.01$, ** $p < 0.05$, * $p < 0.10$.

Table A2: The AI Transparency Paradox: Diagnostic Accuracy by AI Correctness and Explanation Status

	AI Advice is Correct		**AI Advice is Incorrect**	
	No Explanation	Explanation	No Explanation	Explanation
Diagnostic Accuracy	87.4%	93.7%	14.3%	9.4%
Standard Error	(0.6)	(0.4)	(1.2)	(1.0)
Observations	1,419	1,408	516	512
Explanation Effect	+6.0 pp***		-4.4 pp**	
95% CI	[3.6, 8.5]		[-8.6, -0.3]	
Overall Effect	+3.2 pp***			
Net Benefit	73% × 6.0pp - 27% × 4.4pp = +3.2 pp			

Notes: Diagnostic accuracy is the probability assigned to the correct diagnosis. Standard errors in parentheses, clustered at the physician level. Explanation effects show the difference between explanation and no-explanation conditions within each AI correctness category. *** p<0.01, ** p<0.05, * p<0.10.

Table A3: Transparency Paradox Robustness Across AI Advice Formats: Diagnostic Accuracy by AI Correctness, Explanation Status and Physicians' prior Correctness

	Deterministic Format		**Probabilistic Format**	
	AI Correct	AI Incorrect	AI Correct	AI Incorrect
No Explanation	89.0%	10.4%	85.7%	18.2%
	(0.9)	(1.5)	(0.9)	(1.9)
With Explanation	95.6%	7.3%	91.7%	11.6%
	(0.6)	(1.4)	(0.5)	(1.4)
Explanation Effect	+6.6***	-2.9	+4.2***	-7.2***
	(1.2)	(2.4)	(1.6)	(3.3)
Paradox Magnitude	9.7 pp		12.6 pp	
Net Benefit	3.9pp***		1.2pp	
Observations	1,408	512	1,419	516

Notes: Deterministic format presents single recommendation ("AI recommends diagnosis B"). Probabilistic format presents probability distribution ("60% B, 30% A, 10% C"). Paradox magnitude is the absolute difference between beneficial and harmful effects of explanation. The explanation effect remain robust with mixed-effect regressions. See Appendix Table XX for details. Net benefit calculated as weighted average using observed AI accuracy rates. Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A4: Mechanism 3: Effect of Explanations on Confidence by Prior Uncertainty

	Prior SSQ Quartile				Pooled
	Q1 (Uncertain)	Q2	Q3	Q4 (Certain)	All
Panel A: When AI is Correct					
No Explanation	0.803	0.892	0.920	0.953	0.890
With Explanation	0.846***	0.902	0.955***	0.979***	0.922***
Explanation Effect	+0.048	+0.007	+0.036**	+0.027*	+0.029*
Panel B: When AI is Incorrect					
No Explanation	0.781	0.834	0.839	0.912	0.840
With Explanation	0.819	0.837	0.922***	0.957**	0.886***
Explanation Effect	+0.038	+0.002	+0.082***	+0.037	+0.041**
Panel C: Confidence-Accuracy Correlation					
No Explanation	0.24	0.23	0.29	0.22	0.25
With Explanation	0.24	0.29	0.20	0.15	0.23
Observations	975	960	960	960	3,855

Notes: Prior SSQ quartiles measure initial diagnostic uncertainty (Q1 = most uncertain, Q4 = most certain). Panel C shows correlation between confidence (SSQ) and accuracy within each group. The declining correlation with explanations indicates increased false confidence. Mixed-effects regression with participant random effects. *** p<0.01, ** p<0.05, * p<0.10.

B Instructions to Participants

Welcome

Welcome to this decision-making experiment. Please make your decisions carefully, as you will be paid based on the choices you make. Additionally, based on your decisions, a donation will be made to a patient-regarding charity.

The Scenario

You will be presented with **15 medical scenarios** and asked to perform **diagnostic assessments** based on a patient's condition. Each scenario includes **five possible diagnoses** (A, B, C, D, E), with **only one** correct. You will assign probabilities to each option, indicating how likely you believe each is correct.

The AI Advice

You will receive **AI advice** (generated by ChatGPT 4, accuracy $> 70\%$) providing **[one/two] suggested option(s)** [and a short explanation]. The AI's suggestion is not guaranteed to be correct.

Overview of Your Tasks

In each scenario, you will be instructed to complete **three tasks**:

- **Task 1: Initial Estimations** (before seeing AI advice)
- **Task 2: Expected Informativeness Score** (still before AI advice)
- **Task 3: Final Estimations** (after seeing AI advice)

We use a standard quadratic scoring rule to determine your payoffs in each task (you may ask us the details after the experiment), this rule makes sure that being honest and accurate is in your best interest.

Task 1: Initial Estimations (Before AI Advice)

Your Task: Read the scenario describing the patient's condition. Distribute probabilities across the five options (A, B, C, D, E) indicating how likely you believe each is, all probabilities shall sum to 100%.
Payment: At the end of the scenario, the correct diagnosis will be revealed. Your payoff depends on how close your stated probabilities are to the truth, evaluated by a *quadratic scoring rule*. The maximum payment for this task is *2* Yuan.

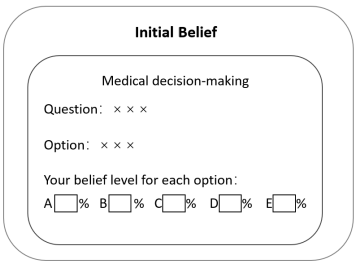

Figure 1: Example Screenshot of Initial Estimation

Task 2: Expected Informativeness Score (Before AI Advice)

Definition: Your *Informativeness Score* measures how certain (or "informative") your probability estimates are. It is calculated as:

$$\sum_{t \in \{A,B,C,D,E\}} (\text{Initial Estimation}_t)^2.$$

An even split of 20% each yields $0.2^2 \times 5 = 0.2$, while being 100% sure of one option yields $1^2 = 1$. Hence, more confident assessments produce higher informativeness scores.

Your Task: Right after giving your initial estimates, you will see your informativeness score. Then, predict how you think *this score will change* after viewing the AI advice.

Payment: You will be rewarded based on how accurate your prediction is. The maximum payment for this is *2* Yuan.

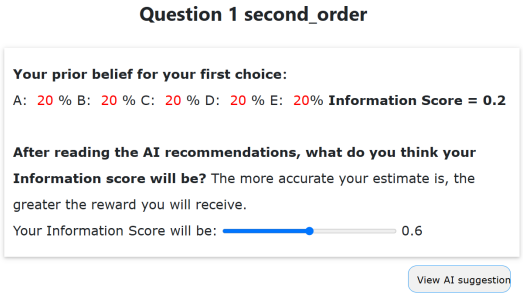

Figure 2: Example Screenshot of Expected Informativeness Score

Task 3: Final Estimations (After AI Advice)

Your Task: After reading the AI's [one/two] suggested option(s) [and explanation], assign probabilities again to the five options (A, B, C, D, E) so that they sum to 100%.

Payment: Your final estimations is also scored using the quadratic scoring rule to make sure it is of your best interest to provide us your actural estimations. The maximum payment for this part is *2* Yuan.

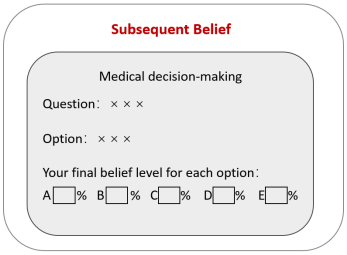

Figure 3: Example Screenshot of Final Estimation

Your Total Payment

Additional Payment: At the start, you will answer **10 multiple-choice questions** (one correct option among five). Each correct answer earns you **1 Yuan**.

Overall Payment: You will receive payoffs for each of the 15 scenarios from: - Task 1 (Initial Estimations) - Task 2 (Expected Informativeness Score) - Task 3 (Final Estimations) plus any additional payments from the 10 initial questions.

Donation to Charity

A donation will be made on your behalf to a patient-focused charity based on your payoffs in the Final Estimation Tasks. Higher earnings lead to a larger donation to help patients with ALS, up to a maximum of 10 Yuan.

Before the experiment begins, you will answer several questions to ensure you understand these instructions. Your responses **will not** affect your payment. If you have any questions, please raise your hand.

Thank you for your participation and good luck!

Post-experimental questionnaire

Questionnaire 1

1. What is your gender?
   ○ Male
   ○ Female
2. What is your age?
   ○ 18-20 years old
   ○ 21-23 years old
   ○ 24-26 years old
   ○ 27-29 years old
   ○ Over 30 years old
3. What is the duration of your education?
   ○ 2 years or less
   ○ 3 years
   ○ 4 years
   ○ 5 years
   ○ More than 5 years
4. What is your highest education level?
   ○ Secondary Specialized
   ○ Associate Degree
   ○ Bachelor's Degree
   ○ Master's Degree
   ○ doctorate
5. If you unexpectedly received 1000 yuan today, how much would you donate to charitable causes?
6. How willing are you to donate to public welfare causes without expecting anything in return?
   ○ Extremely unlikely
   ○ Unlikely
   ○ Neutral
   ○ Likely
   ○ Extremely likely
7. I believe people are generally well-intentioned.
   ○ Extremely unlikely
   ○ Unlikely
   ○ Neutral
   ○ Likely
   ○ Extremely likely

Questionnaire 2: Raven test

This test task contains 6 figures as shown below example, each figure consists of 3x3 different patterns,where the last pattern is blank. Each row/column of these patterns is arranged

according to certain rules. Your task is to find the best pattern to fill in the blanks. You have a total of 5 minutes to complete this part of the test. For every correct answer, you will get an extra 1 yuan.

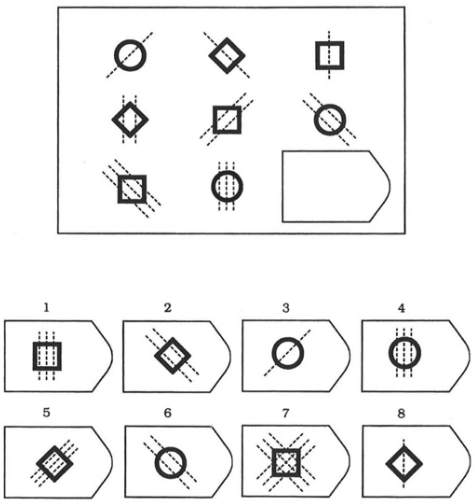

Figure A1: An example of a Raven task test

Questionnaire 3: CRT test

Please answer the following four questions. You have a total of 3 minutes to complete this part of the test. You will receive an additional 1 yuan bonus for each correct answer of these questions.

1. A pair of tennis rackets and a ball together cost 1.10, and the racket is 1 more expensive than the ball. What is the price of the ball in dollars?

2. If 5 machines can produce 5 parts in 5 minutes, how many minutes would it take for 100 machines to produce 100 parts?

3. A barrel of pure water would be finished by Xiao Ming in 6 days and by Xiao Hong in 12 days. If Xiao Ming and Xiao Hong become roommates and drink from the same barrel, how many days would it take for them to finish the water?

4. As shown in the figure below (which is not provided here), there are four cards (A, B, C, D) on the table. Each card has a number on the front and a color on the back. Now, Xiao Ming has made the following conjecture: If the front of a card is an even number, then its back is blue. Assuming you can look at these cards, which cards must you turn over to verify whether Xiao Ming's conjecture is correct?

Questionnaire 4: Algorithm literacy and awareness

Below are some descriptions related to the field of artificial intelligence. For each statement, please select according to your thoughts.

1. Are you aware of the current state of development in artificial intelligence (AI)?

○ Completely Unaware ○ Not Very Aware ○ Somewhat Aware ○ Aware ○ Very Aware

2. Are you aware of the potential use of AI algorithms in medical decision-making?

○ Completely Unaware ○ Not Very Aware ○ Somewhat Aware ○ Aware ○ Very Aware

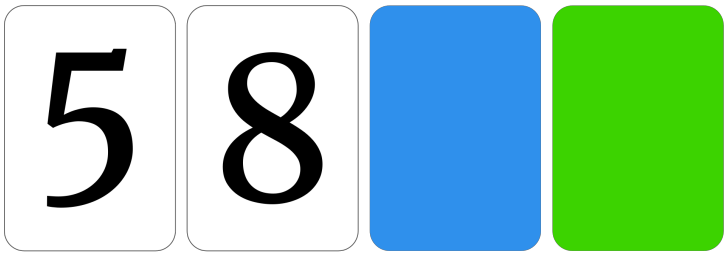

Figure A2: An example of a CRT test

3. Do you think AI algorithms will have what kind of impact on the future development of the medical industry?

○ Very Negative Impact ○ Negative Impact ○ No Impact ○ Positive Impact ○ Very Positive Impact

4. Are you aware of the basic principles of machine learning?

○ Completely Unaware ○ Not Very Aware ○ Somewhat Aware ○ Aware ○ Very Aware

5. Do you believe that a critical thinking approach should be maintained when using AI algorithms?

○ Strongly Disagree ○ Disagree ○ Neutral ○ Agree ○ Strongly Agree

Below are some relevant descriptions of AI for the healthcare field, for each of the following statements, please choose as you see fit.

1. How useful do you think AI algorithms are in medical decision making?

○ Almost no effect ○ Limited role ○ Somewhat useful ○ Very useful

2. How useful do you think AI algorithms are in assisting physicians' treatment systems?

○ Almost no effect ○ Limited role ○ Somewhat useful ○ Very useful

3. How useful do you think AI algorithms are in healthcare data analysis?

○ Almost no effect ○ Limited role ○ Somewhat useful ○ Very useful

4. Have you ever questioned or validated the results of AI algorithms in healthcare decision-making?

○ Not at all ○ Only once ○ A few times ○ Regularly

5. Would you be willing to receive specialized training on the use of AI in healthcare?

○ Totally unwilling ○ Not very willing ○ Willing ○ Very willing

Questionnaire 5: Algorithm fairness and trust

1. Which AI large language model have you used?

□ ChatGPT
□ Gemini
□ Wen Xin Yi Yan
□ Kimi chat
□ Sora
□ Dall.E3
□ Other

□ Don't know

If you selected "Other", please enter your answer below.

2. How often do you use the Large Language Model?

○ Every day ○ 3 times a week ○ Once every six months ○ Never

3. Below are two different perspectives on Al for healthcare, for each of the following statements, please choose as you see fit. Whichever side of the argument you agree with, select the scale point that is close to that argument.(There are five points on the scale, and the meanings from left to right are "Strongly Agree With Viewpoint A", "Somewhat Agree With Viewpoint A", "Neutral", "Strongly Agree with Viewpoint B", "Strongly Agree with Viewpoint B").

(1)View A : AI technology has made it more difficult for patients in low-income or remote areas to access quality healthcare.

View B : AI technology makes it more likely that patients in low-income or remote areas will have access to high-quality healthcare.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(2)View A : AI technology has a significant impact on healthcare access equity.

View B : AI technology did not have any significant impact on equity of access to healthcare.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(3)View A : AI technology makes resources tend to be0allocated to patients or providers who can pay higher fees.

View B : AI technology has led to a more even distribution of resources.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(4)View A : The use of AI technology may exacerbate inequalities in research and development of effective treatments for certain diseases.

View B : AI technology has made it possible for rare diseases to be more fully researched as well.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(5)View A : AI technology may cause physicians to become overly reliant on technology at the expense of direct patient interaction.

View B : AI technology may increase patient trust in healthcare because it provides more accurate medical information.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(6)View A : AI provides inaccurate results.

View B : AI provides accurate results.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(7)View A : AI provides results that are not easily applied to common problems.

View B : AI provides enough results to apply to common problems.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(8)View A : If hospitals introduce AI-assisted medical technology, I'm not so sure about using AI for assisted diagnosis and treatment.

View B : If the hospital introduces AI-assisted healthcare technology, I will combine it with a large model to assist in the consultation.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(9)View A : Using the Large Language Model will not improve the accuracy of a physician's diagnosis.

View B : Using the Large Language Model can improve the accuracy of a physician's diagnosis.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B

(10)View A : AI algorithms have a very negative impact on the future of the healthcare industry.

View B : AI algorithms have a very positive impact on the future of the healthcare industry.

Strongly Agree With A ○—-○—-○—-○—-○ Strongly Agree With B