When Life Gives You AI, Will You Turn It Into A Market for Lemons? Understanding How Information Asymmetries About AI System Capabilities Affect Market Outcomes and Adoption

Alexander Erlei
University of Göttingen
Göttingen, Germany
alexander.erlei@wiwi.uni-goettingen.de

Federico Maria Cau
Department of Mathematics and Computer Science
University of Cagliari
Cagliari, Italy
federicom.cau@unica.it

Radoslav Georgiev
Delft University of Technology
Delft, Netherlands
r.k.georgiev@student.tudelft.nl

Sagar Chethan Kumar
Columbia University
New York, New York, USA
sagar.chethankumar@columbia.edu

Kilian Bizer
University of Goettingen
Goettingen, Germany
bizer@wiwi.uni-goettingen.de

Ujwal Gadiraju
Web Information Systems
Delft University of Technology
Delft, Netherlands
u.k.gadiraju@tudelft.nl

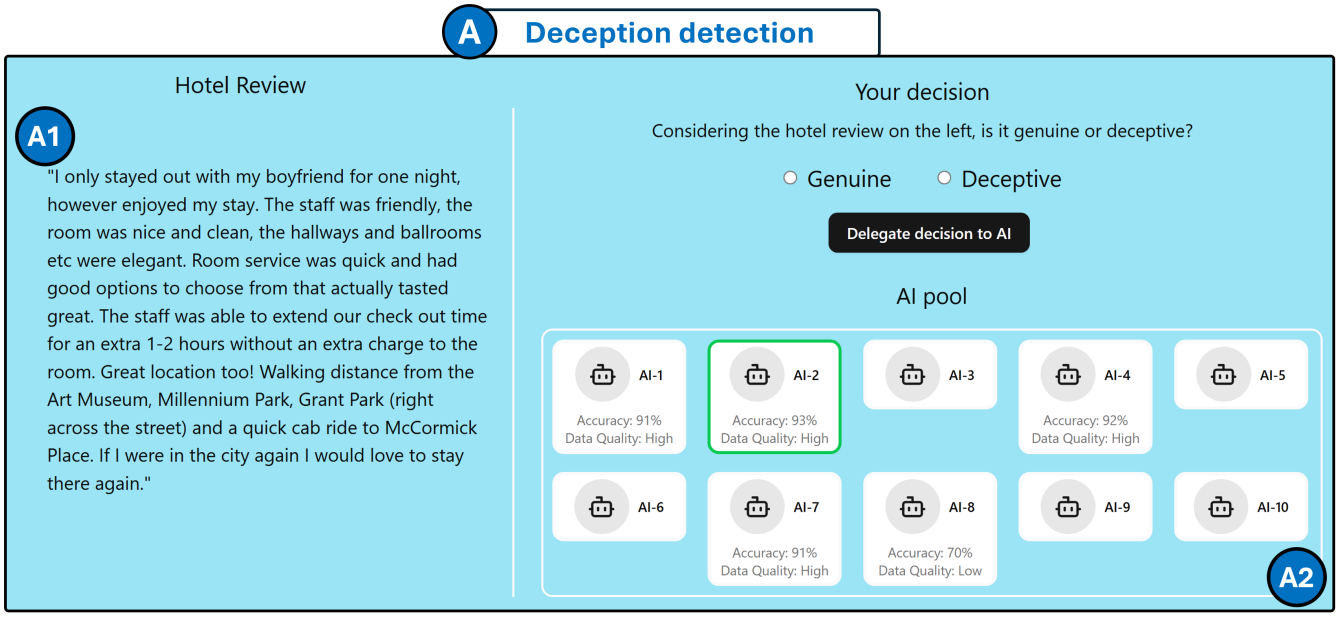

Figure 1: Task interface that participants interacted with during the study. This illustration represents the *deceptive review detection* task (A) with full information disclosure about the AI systems in the available pool. On the left side (A1), participants viewed the hotel review they needed to verify. On the right side, participants saw the binary decision options for the current trial, and an AI pool of ten AI systems (A2). By hovering over an AI system, participants could access the available information (*i.e.,* quality indicators) corresponding to the AI system in conditions with partial or full information disclosure. Participants could select an AI system from the pool for delegation by first clicking on it—which makes the selected AI system light up with a green border—and then clicking the '*Delegate decision to AI*' button to complete the current trial.

CHI '26, Barcelona, Spain

ACM ISBN 979-8-4007-2278-3/26/04
https://doi.org/10.1145/3772318.3791420

Abstract

AI consumer markets are characterized by severe buyer-supplier market asymmetries. Complex AI systems can appear highly accurate while making costly errors or embedding hidden defects. While there have been regulatory efforts surrounding different forms of disclosure, large information gaps remain. This paper provides the

first experimental evidence on the important role of information asymmetries and disclosure designs in shaping user adoption of AI systems. We systematically vary the density of low-quality AI systems and the depth of disclosure requirements in a simulated AI product market to gauge how people react to the risk of accidentally relying on a low-quality AI system. Then, we compare participants' choices to a rational Bayesian model, analyzing the degree to which partial information disclosure can improve AI adoption. Our results underscore the deleterious effects of information asymmetries on AI adoption, but also highlight the potential of partial disclosure designs to improve the overall efficiency of human decision-making.

CCS Concepts

• **Human-centered computing** → **Human computer interaction (HCI)**; • **Information systems**; • **Computing methodologies** → *Artificial intelligence*;

Keywords

Human-AI interaction, AI adoption, information asymmetry, human-AI decision-making

ACM Reference Format:

Alexander Erlei, Federico Maria Cau, Radoslav Georgiev, Sagar Chethan Kumar, Kilian Bizer, and Ujwal Gadiraju. 2026. When Life Gives You AI, Will You Turn It Into A Market for Lemons? Understanding How Information Asymmetries About AI System Capabilities Affect Market Outcomes and Adoption. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems (CHI '26), April 13–17, 2026, Barcelona, Spain.* ACM, New York, NY, USA, 24 pages. https://doi.org/10.1145/3772318.3791420

1 Introduction

AI systems are increasingly integrated into consumer-facing applications, yet their complexity often obscures critical quality dimensions such as reliability, safety, and fairness from end-users [68, 108, 109]. While these systems can appear highly accurate, they may still make costly mistakes or exhibit hidden defects [29, 38, 90, 116, 131, 135, 143]. From an HCI perspective, this opacity creates significant challenges for trust calibration, appropriate reliance, and user agency. Users often lack the ability to verify system quality either before or after use [18, 49], which can lead to misaligned mental models and poor decision-making.

In the field of economics, this is traditionally framed as a "market for lemons" problem, where uncertainty about product quality drives inefficient outcomes [2]. In the context of human-AI interaction, this translates into information asymmetry between designers or deployers of AI systems and their users, where users cannot easily assess whether an AI system is fit for their task. This can lead to sub-optimal user decisions on whether to adopt an AI system, or which AI system to adopt, given the presence of alternatives. While mechanisms such as reputation systems, third-party certifications, and disclosure labels (e.g., Apple's privacy "nutrition" labels) aim to bridge this gap [60, 126, 127, 132], they often fail in practice due to selective reporting, high cognitive load, or strategic complexity that undermines transparency [3, 73, 91]. Recent regulatory efforts, such as the EU AI Act, mandate transparency, but standards remain fragmented and lack empirical grounding in user behavior [77, 142]. Evidence suggests that current disclosure practices (e.g., model cards) vary widely in informativeness and usability [91, 120]. Importantly, providers of these systems are generally incentivized to reveal favorable but hide unfavorable information [72, 117], or use unnecessary complexity to shroud information [73], which constraints the usefulness of voluntary transparency efforts. Addressing this requires a regulatory framework that is strategy-proof, such that suppliers cannot "game" the system through partial information revelation [114].

In this paper, we provide the first experimental evidence on how information asymmetries (modeled through disclosure design) and the relative density of low-quality AI systems (also referred to as *lemons*) in the available pool for users, influence the adoption and reliance on AI systems. To this end, we simulate an interactive decision-making environment in which participants complete three different tasks across 10 rounds, in each of which they can decide to use assistance from an AI system of their choice from a pool of alternatives. Our controlled study follows a 3 (lemon density: **`Low Density`** vs **`Medium Density`** vs **`High Density`**) × 2 (information disclosure: **`No Disclosure`** vs **`Partial Disclosure`**) between-subjects design, with an additional benchmark condition of **`High Density`** with **`Full Disclosure`**. The overall quality of an AI system in our study is represented by an *accuracy score* and a *data quality* score, where the latter represents the AI system's generalizability.[1] In the **`Partial Disclosure`** conditions, participants only observe the accuracy score, whereas the **`Full Disclosure`** conditions reveal both accuracy and data quality. This design allows us to examine how users interpret and act on these cues under varying conditions of low-quality AI system density, as illustrated in Fig. 1. We address the following research questions:

- **RQ1**: How do information asymmetries about AI system capabilities affect users' adoption of AI and market outcomes in the presence of low-quality AI systems?
- **RQ2**: How do different information disclosure requirements about AI system capabilities impact user behavior, market outcomes, and reliance on low-quality AI systems?

We contribute a novel experimental framework that adapts the classic economics theory of "market for lemons" [2] to human-AI interaction, operationalizing lemon density and disclosure strategies to systematically study how information asymmetry and uncertainty shape user reliance and decision-making. Our findings show that users are sensitive to the presence of low-quality AI systems when no disclosure is provided (**`No Disclosure`**), but they struggle to calibrate their reliance effectively over time. When the proportion of low-quality systems is low, participants tend to underutilize AI assistance, missing opportunities to improve performance. Conversely, when low-quality systems are prevalent, they over-rely

[1] Note that our technical operationalization of AI system quality does not consider other relevant dimensions of system quality such as robustness, fairness, uncertainty, or alignment with human values—where data quality remains a fundamental enabler. This operationalization is a principled simplification that necessarily excludes other important dimensions of AI system quality that fall outside the scope of our study, while remaining externally valid. Several dimensions of AI system quality in the real world are often encapsulated in composite metrics that are conveyed to consumers to help shape their mental models of the systems.

on AI, delegating tasks even when it is detrimental. While some learning occurs across rounds, its magnitude is small, suggesting persistent challenges in forming accurate mental models of system quality.

We found that introducing partial disclosure (i.e., revealing only AI system accuracy) significantly improves decision-making efficiency. Participants leverage these incomplete signals to avoid low-quality systems at high rates, consistent with adaptive trust calibration. This leads to substantial performance gains in **`Low Density`** and **`Medium Density`** conditions, though benefits diminish when low-quality systems dominate (i.e., in the **`High Density`** conditions). The participants' ability to interpret and act on partial disclosure cues declines as the density of poor-quality systems increases, highlighting the limits of simple transparency mechanisms. Importantly, partial disclosure does not change overall delegation rates or the proportion of participants who fail to optimize their choices, indicating that information disclosure primarily improves the quality of delegated decisions rather than the quantity. This effect is strong enough to offset a doubling of low-quality systems in the market. In contrast, our full disclosure benchmark (**`Full Disclosure`** with **`High Density`**)–where users see both accuracy and data quality–reveals a different challenge: even when fully informed, participants exhibit inefficiently low AI adoption. While they successfully avoid low-quality AI systems, only about 58% of predictions are delegated to systems that are correct with a probability of 90%, resulting in average losses of around 20% compared to full delegation. This suggests that a greater degree of information disclosure does not automatically translate into better outcomes, and that other factors (e.g., risk aversion) may lead users to under-utilize AI systems even when it is objectively beneficial. We also found that the average performance in **`Partial Disclosure`** with **`Low Density`** matches that of **`Full Disclosure`** with **`High Density`**, underscoring the importance of designing disclosure for actionability rather than completeness.

While disclosure mitigates information asymmetry, it does not eliminate other cognitive and behavioral barriers such as bounded rationality [78], risk aversion, and trust calibration [98] challenges. This nuance shows that system transparency is necessary but not sufficient to foster appropriate adoption and reliance on AI systems. Apart from important design implications for human-AI interaction, our work has direct implications for emerging standards such as the EU AI Act, where enforceable yet lightweight disclosure rules may be more effective than complex, fully detailed reporting in promoting appropriate reliance on AI systems.

2 Background and Related Work

2.1 Human Beliefs and Reliance on AI Systems

Behaviorally, our paper relates to the literature at the intersection of user information, user beliefs, and AI utilization. Biermann et al. [18] provide experimental evidence that people struggle ex ante and ex post to verify the prediction quality of algorithms, which can even be exacerbated by explanations. Users learn over time, which has implications for small but not for large markets. In general, humans use accuracy signals to update their beliefs and adjust reliance behavior, but tend to calibrate imperfectly [21, 47, 63]. Furthermore, additional information may also fail to improve human reliance if they cannot interpret the signals correctly or experience information overload [112]. Research on over-reliance has argued that confirmation bias, anchoring bias and base neglect may explain why people sometimes apparently under-weigh observable shortcomings of AI decision making systems [54, 99, 104, 106, 115]. In the context of explainable AI (XAI), users have exhibited a variety of cognitive biases related to the processing of information signals, such as representative and availability biases, choice overload, or relative inattention to false positives [14]. In our experiment, we rely on information labels as partial signals that allow rational users to improve their decision-making but cannot rule out false negatives (i.e. accidentally approaching a lemon). False positives (wrongly identifying a high-quality AI system as a lemon) do not play a role [see also 58, 82]. The main reason is that belief updating under information asymmetries across different tasks is already cognitively demanding. Here, focusing on false negatives is not only externally more valid for regulatory purposes, but also streamlines the cognitive effort of participants to the processes relevant to our main research questions. Agarwal et al. [1] analyze how radiologists with access to AI assistants update their beliefs and concurrent diagnostic choices when the AI signal is (i) certain or (ii) uncertain. They find improvements in decision quality under certainty, while radiologists fail to update under uncertainty. In Reverberi et al. [119], endoscopists were able to rationally integrate AI advice into their diagnostic process. Note that in these cases, experts interact with expert systems with verified, although on a case-by-case basis uncertain, prediction quality. Thus, there is individual-level evidence that (expert) users can rationally integrate AI-generated advice following Bayesian principles, which may be inhibited by uncertainty in the choice environment.

Researchers have studied how user mental models of AI systems, their experiences, expectations, and various AI metaphors can influence and shape experiential and performance outcomes in human-AI collaboration and decision-making contexts [11, 67, 81]. Recently, Gadiraju and Balayn [54] synthesized various human, task, and AI system factors empirically shown to shape outcomes in human-AI collaboration. Prior HCI work on trust calibration and transparency has largely focused on static disclosure artifacts (e.g., model cards [103], impact assessment reports [22]) or interpretability cues in isolation. Our work introduces a behavioral economics lens to HCI, operationalizing the "market for lemons" theory as an experimental framework to study AI adoption under varying information asymmetries and varying densities of low-quality AI systems. This is a novel theoretical integration that connects economic models of uncertainty with human-centered design challenges.

2.2 Information Asymmetry in Human-AI Collaboration: Technical Solutions and Regulatory Frameworks

The asymmetry of information between AI systems and end-users is particularly evident in black-box models, whose decision-making processes are often opaque and difficult to interpret [93]. This lack of transparency can undermine trust, reduce accountability, and hinder users' ability to evaluate the reliability of AI outputs [25, 146]. Explainable AI (XAI) approaches aim to address this gap by providing intelligible explanations that help users understand the

factors driving model predictions [4, 42, 61, 102]. XAI literature spans different approaches that have been proposed, ranging from local [24, 33, 84, 85] to hybrid [7, 15–17, 30, 138] explanations, which involve the integration of data-centric explanations with model-centric ones to promote improved human-AI decision-making and appropriate reliance.

An alternative approach is the use of inherently interpretable, or "glass-box" models [121]. While black-box models often achieve higher predictive accuracy, glass-box models enhance transparency, enabling users to detect and correct data quality issues such as biases, errors, or mislabeling [34, 121]. High-quality, well-labeled, and representative datasets improve the performance of both model types, but the interpretability of glass-box models makes it easier to identify and address problems in the data [34, 146]. Hybrid approaches that combine accurate black-box models with XAI techniques can offer a balanced solution, maintaining predictive performance while providing users with interpretable insights, which could reduce information asymmetry. In this paper, we represent this real-world context where users can interact with AI systems with varying access to information about indicators of their quality.

Addressing the growing concerns and downstream consequences of information asymmetry in human-AI collaboration, regulatory frameworks in the European Union have been designed to enforce transparency and accountability. The EU AI Act introduces risk-based obligations, requiring providers to disclose when exactly users are interacting with AI systems (EU AI Act, Article 50).[2] It mandates the explicit labeling of AI-generated content and encourages explainability in high-risk systems. As a complement to this, the General Data Protection Regulation (GDPR) ensures that individuals are informed about how their personal data is processed by AI, granting rights to explanation and human intervention in automated decision-making (GDPR, Articles 13–22).[3] In contrast, the Digital Services Act (DSA)[4] targets online platforms, requiring transparency in algorithmic content moderation and recommendation systems, and obligating platforms to disclose AI involvement in personalized content and ads. These regulations collectively contribute to the attempts at reducing information asymmetry in human-AI collaboration and aim to promote responsible AI deployment across different domains [10, 35]. Despite these efforts, complexity of the global AI market and benchmarking practices have meant that users continue to interact with and rely on AI systems with limited access to information about their quality. Our work contributes to understanding the impact of regulatory institutions on AI adoption, user behavior, and market outcomes—particularly when low-quality AI systems (*lemons*) are prevalent in the market. Our experimental framework is both novel and extensible. Future studies can vary other factors such as task complexity, stakes, fairness signals, or robustness indicators using the same market-based paradigm.

2.3 Disclosure

Most research surrounding the desirableness and efficacy of disclosure institutions comes from economics, following the classic "unraveling" logic. When quality is verifiable and disclosure is costless, suppliers generally voluntarily reveal quality (except the worst anti-social types) [60, 100]. In practice, unraveling is frequently incomplete. Sellers face strategic incentives to disclose selectively, delay or bury unfavorable information, or increase complexity to shroud key attributes [46, 53, 72, 73, 117]. These frictions are central to AI markets where suppliers can highlight favorable benchmarks while under-reporting limitations, safety risks, or domain shift – precisely the conditions our design emulates with partially informative labels.

A substantial body of empirical literature shows that third-party verification and mandatory disclosure can (partially) discipline markets with hidden qualities. In restaurants, making hygiene report cards salient improves practices and shifts demand toward higher-grade establishments [70, 71]. In healthcare, public "report cards" affect provider behavior and consumer choice, but also induced risk selection, highlighting design trade-offs, such as the scope of suppliers to "game" the system [43]. Publicized plan ratings change enrollment in health insurance markets [74]. In consumer goods, voluntary disclosure regarding nutrition labeling is asymmetric (good types disclose, bad types often do not), whereas mandatory labeling shifts demand and improves welfare by making negative attributes visible [23, 66, 97]. Broadly, the evidence points towards a positive effect of well-designed disclosure and certification institutions on selection and market-wide product quality. However, there is a lot of heterogeneity, specifically with regards to signal type, credibility, domain context and scope. Most evidence comes from experience goods that are comparatively simple to regulate.

Disclosure's efficacy also depends on how information is presented and processed. For example, salience and simplicity can determine whether consumers are able to exploit signals [23, 36]. Particularly in the AI context under market information asymmetries, simplicity is a complex problem, given the Bayesian nature of the involved information and learning process. Certification experiments show sizable premia for verified quality and reductions in "lemons" risk when credible third parties attest to product characteristics [69]. When markets rely solely on reputational mechanisms, disclosure may be too noisy or manipulable to separate types in thick, anonymous environments [105]. This directly maps to AI markets, where voluntary "model cards" or benchmark disclosures can be selectively curated, and where strategic non-disclosure or obfuscation is profitable in the absence of enforceable, strategy-proof *and* behaviorally valid rules [72, 73, 117].

This study leverages these insights in three ways. First, we test a **`Partial Disclosure`** regime that is realistically noisy. It represents the empirically observed gap between theoretical unraveling and real-world selective disclosure. Second, we benchmark against **`Full Disclosure`** to isolate the upper bound on efficiency when information asymmetries have been solved. Third, we vary the complexity of the market by manipulating the density of lemons, which also allows us to make sharp behavioral predictions.

3 Method

This section outlines how we developed the user study to assess the "market for lemons" problem for AI adoption, starting with an overview of the lemon market we simulate in our study, the

[2] https://artificialintelligenceact.eu/article/50/ [last accessed on: 11 September 2025]
[3] https://gdpr-info.eu/art-13-gdpr/ [last accessed on: 11 September 2025]
[4] https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package [last accessed on: 11 September 2025]

selection of different types of tasks, the AI system pool, instances, and the design of the AI-assisted interface.

3.1 Simulation of The Lemon Market

Our study aims to simulate a market environment in which consumers demand some AI good (e.g., an AI system that can aid a decision-making task) but are ex ante uncertain about the system's quality. Consumers observe the quality after usage, but operate in a large market, such that their experiences do not meaningfully affect their search set, and hence do not reduce uncertainty for future purchases. There is a single posted price because low-quality sellers mimic high-quality ones. Our simulation of the lemon market is motivated by the increasing reliance of users on AI systems to support decisions today (e.g., choosing a resume screener, a grammar checking tool, or a writing aid). Yet users often cannot tell whether an AI system is broadly reliable before they commit to it. Benchmark accuracy labels can look impressive, but may not translate to real-world generalizability and mislead users (e.g., due to biased data or domain shifts). Thus, simulating this market can allow us to systematically study how different levels of disclosure shape user adoption of AI systems in a market where some AI systems are truly high quality (i.e., *peaches*) and others are low quality (i.e., *lemons*).

Formally, there are two types of AI systems $\theta \in \{H, L\}$ characterized by quality $Q = (\alpha, g)$ that differ in terms of measurable accuracy $\alpha \in [0, 1]$ and generalizability due to data quality $g \in [0, 1]$. High-quality AI systems (i.e., *peaches*) H always exhibit $Q_H = (\alpha_H, g_H)$, and low-quality AI systems (i.e., *lemons*) exhibit either $Q_L = (\alpha_H, g_L)$ or $Q_L = (\alpha_L, g_L)$ with $g_H > g_L$, and $\alpha_H > \alpha_L$. Here, we assume that lemons always exhibit poor generalizability (g_L) due to fundamental data issues such as dataset biases or sampling issues, overfitting to training distributions, or domain shifts grounded in real-world considerations. However, some lemons can exhibit high accuracy (α_H) in external tests. For lemons, there is heterogeneity in measurable AI accuracy from external testing, e.g., through benchmark platforms. Thus, lemons exhibiting high accuracy scores have incentives to disclose these scores voluntarily. Peaches, on the other hand, always exhibit high accuracy scores and are not plagued by data-related issues. Therefore, accuracy is only partially informative. Consumers earn some gross surplus $u_H > 0$ from using a high-quality AI system, but experience a cost $u_L < 0$ when purchasing a low-quality AI system.[5] The uniform market price is p. Depending on the experimental conditions, consumers can (not) observe the share of lemons in the market $\lambda \in [0, 1]$ and receive a signal s about the quality of each AI system based on three information regimes. In the **`No Disclosure`** conditions, consumers condition their choices based on what they learn about AI system quality in the market across multiple trials. In the **`Partial Disclosure`** conditions, consumers observe a signal $s \in 0, 1$ in the form of an accuracy badge label that represents the measurable accuracy (either α_H or α_L). If $s = 1$, the product has high accuracy. Consumers always know that there is a probability $\gamma = 1$[6] for $s = 1$ if $\theta = \mathbf{H}$, and a probability $0 < \beta < \gamma$ for $s = 1$ if $\theta = \mathbf{L}$. Formally, $\mathbf{Pr}[s = 1|H] = \gamma$, $\mathbf{Pr}[s = 1|L] = \beta$. Then, consumers who observe the signal s update their beliefs that any AI is of high quality using Bayes' rule after seeing the accuracy label:

$$\mathbf{Pr}[\theta = H|s = 1] = \frac{(1-\lambda)\gamma}{(1-\lambda)\gamma + \lambda\beta},$$

$$\vee \quad \mathbf{Pr}[\theta = H|s = 0] = \frac{(1-\lambda)(1-\gamma)}{(1-\lambda)(1-\gamma) + \lambda(1-\beta)}.$$

A key feature of our signaling structure is that while the $s = 1$ signal is noisy, the $s = 0$ signal is perfectly informative and always reveals a lemon. This simplifies exposure for consumers and gives a sharp behavioral prediction whereby rational consumers never purchase a low-accuracy AI product. It is the minimal requirement for partial disclosure regulations. Finally, the **full** disclosure reveals θ perfectly, consumers are fully informed. To elucidate, even a simple badge (that supports partial disclosure) can reduce harmful AI adoption if low-accuracy labels reliably flag lemons. But unless generalizability is disclosed, some lemons will still pass as high accuracy. Full disclosure solves adverse selection by aligning user decisions with true quality.

We make five important assumptions. One, consumers are risk-neutral and maximize expected monetary surplus $E[u_\theta] - p$. Two, all agents know $\lambda, u_H, u_L, p, \gamma, \beta$.[7] Three, the search space is constant across rounds, each purchase is independent (large market). Fourth, lemon sellers post the same price as high-quality sellers. This rules out, for example, side-payments, or incentives for endogenous segmentation. Fifth, disclosure changes neither production cost, nor prices in the short run, p is fixed throughout the experiment.

Prediction. The risk-neutral consumer buys if $EU(x) = Pr[\theta = H|x]u_H + Pr[\theta = L|x]u_L - p \geq 0$ where $x \in \{\emptyset, s = 1, s = 0, \theta = H, \theta = L\}$. In **`No Disclosure`**, they can only condition their decision on the (learned) prior λ: $EU(\emptyset) = (1-\lambda)u_H + \lambda u_L - p$. In **`Partial Disclosure`**, $EU(s = 1) = [\frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_H + [1 - \frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_L - p$ or $EU(s = 0) = [\frac{(1-\lambda)(1-\gamma)}{(1-\lambda)(1-\gamma)+\lambda(1-\beta)}]u_H + [1 - \frac{(1-\lambda)(1-\gamma)}{(1-\lambda)(1-\gamma)+\lambda(1-\beta)}]u_L - p$. In **`Full Disclosure`**, we get $EU(H) = u_H - p > EU(L) = u_L - p$. Finally, we assume $p < u_H$ such that it makes sense for consumers to enter the market for AI products. Using this simple decision model, we set our experimental parameters such that, on average given priorly published behavioral data from user studies (see section Experimental Parameters below), $(1-\lambda)u_H + \lambda u_L > p$ in **`Low Density`**, $(1-\lambda)u_H + \lambda u_L < p < [\frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_H + [1 - \frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_L$ in **`Medium Density`**, and $[\frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_H + [1 - \frac{(1-\lambda)\gamma}{(1-\lambda)\gamma+\lambda\beta}]u_L < p < u_H$ in **`High Density`**. Note that we focus on a short-run analysis with exogenous prices. Our main focus is on consumer behavior, and the effectiveness of different disclosure regimes, rather than seller behavior. If price were endogenous in a competitive market with free entry, high quality sellers need $p \geq c_H$, and lemons need $p \geq c_L$ where $c_H > c_L$. Consumers cannot separate types, and are therefore willing to pay up to $p = E(x)$. For instance, in **`No Disclosure`**, their willingness

[5] Consumer surplus in economics is the monetary gain that consumers make when they purchase a product or service for a price that is less than the highest price they are willing to pay [96].

[6] Results are unchanged for $\gamma < 1$ as long as $\gamma > \beta$.

[7] In **`Partial Disclosure`** and **`Full Disclosure`**, λ is observable. In **`No Disclosure`**, subjects cannot observe λ, but learn it over the course of 30 rounds. Here, the prediction hinges on consumer (i) exploration of the AI market and (b) learning.

to pay (WTP)[8] is $E(\emptyset) = (1-\lambda)u_H + \lambda u_L$. As λ rises, $E(x)$ and hence the market price p falls, where eventually $p < c_H$ and high-quality sellers exit the market, once again increasing λ, leading to the well-known "death spiral" [2]. Full disclosure solves that problem by driving lemons prices downwards to $p = c_L$.

3.2 Experimental Parameters

In our experiment, participants choose between themselves and an AI system. Instead of setting a fixed exogenous price, we model the price as the average opportunity costs of relying on an AI system, i.e., average subject performance. Prior research suggests user accuracy of 50%–60% for all three chosen tasks [26, 28, 63, 110, 113]. Following that, we assume an average prediction accuracy of 55%. We set average lemon accuracy to 15% and average accuracy of a high-quality AI system to 90%. Participants earn 30 Coins ($0.1) per correct prediction, and 0 otherwise. This gives $p = 16.5$, $u_H = 27$ and $u_L = 4.5$. The publicly known probability of a lemon to exhibit high accuracy is $\beta = \frac{1}{3}$. The risk-neutral expected payoffs (assuming that participants correctly use signal s and, in **`No Disclosure`**, learn the prior λ) are shown in the Table 1 below.

Table 1: Expected Average AI-Market Payoffs Between Conditions

	`Low Density`	`Medium Density`	`High Density`
`No Disclosure`	4.5	-3	-8.25
`Partial Disclosure`	8.79	3	-4.875
`Full Disclosure`	—	—	12

The expected payoff depends on both density and disclosure institution, where under **`Medium Density`**, partial disclosure flips the expected payoff from negative to positive. In **`High Density`**, relying on the AI market is economically harmful in expectation, except if information asymmetries are fully resolved. Our predictions follow these payoffs. Higher expected returns increase the share of subjects who – if they efficiently utilize the information signals – rely on the AI market. Hence, delegation increases along the disclosure institutions and decreases along the density condition. The majority of subjects consistently use the AI market in **`Low Density`** irrespective of the disclosure condition, whereas in **`Medium Density`**, this pattern only emerges in **`Partial Disclosure`**. For **`High Density`** we predict very low rates of AI adoption, with higher delegation shares in **`Partial Disclosure`** than **`No Disclosure`**.

3.3 Hypotheses

Following the conditions described above, we derive four hypotheses that we aim to test in our study:

- **H1** (*effect of disclosure*): AI adoption will increase with the level of disclosure (i.e., AI adoption in the experimental conditions will follow: **`Full Disclosure`** > **`Partial Disclosure`** > **`No Disclosure`**.)
- **H2** (*effect of lemon density*): AI adoption will decrease as the share of lemons in the market increases (i.e., AI adoption in the experimental conditions will follow: **`Low Density`** > **`Medium Density`** > **`High Density`**.)
- **H3** (*interaction*): AI adoption will vary across the different lemon density and disclosure conditions as follows:

H3a: In **`Low Density`** conditions, participants will consistently delegate to AI across all disclosure conditions.

H3b: In **`Medium Density`** conditions, participants will delegate only under **`Partial Disclosure`** or **`Full Disclosure`**.

H3c: In **`High Density`** conditions, participants will delegate only under **`Full Disclosure`**.

- **H4** (*decision-making efficiency*): When disclosure is available, participants will use the information efficiently (e.g., avoid low-accuracy badges), leading to higher decision-efficiency compared to **`No Disclosure`**.

3.4 Tasks Selection

To gain a broader understanding of how information asymmetry and disclosure impact AI adoption in a lemon market, we selected three different tasks with varying data modality that are widely used in the Human-Centered AI (HCAI) literature: skin cancer prediction, loan approval, and deceptive review detection. Note that all task data are appropriately anonymized and available publicly.

3.4.1 Skin cancer prediction. For image data, we opted for the *ISIC 2018 challenge dataset*[9] given its suitability for AI-assisted decisions as in previous work [12, 28, 32, 139, 140, 145]. Specifically, we set up a binary skin cancer prediction task where participants need to decide, given a skin image, whether it appears as benign or shows signs of malignant skin cancer (e.g., a general skin cancer).

3.4.2 Loan approval. For tabular data, we selected the *loan prediction problem dataset*[10] as a testbed, given its utility in several previous works in AI-assisted decisions [19, 39, 48, 57, 59, 61, 63, 141]. In this task, participants are asked to decide whether to accept or reject a loan application based on twelve attributes of an applicant (e.g., employment status, education level, credit history, etc.). To avoid creating an ambiguity effect as shown in prior literature [44], we removed the Loan-ID attribute, as it provides little information for decision-making—resulting in eleven attributes.

3.4.3 Deceptive review detection. For text data, we chose *deceptive hotel reviews dataset*[11] since it has been used in previous studies of human-AI collaboration [8, 26, 62, 83, 110, 113]. The goal of the task is to determine whether a review is "genuine", hence written by someone who stayed at the hotel, or "deceptive", so written by someone who has not. As done in previous work [26], we selected genuine reviews from online sites such as TripAdvisor that involve positive polarity. Instead, deceptive reviews were collected from Amazon Mechanical Turk workers.

3.5 Instances and AI System Selection

For each task, we selected ten instances, ensuring a balance between the binary true classes, with five for the positive class (i.e.,

[8] WTP is a core concept in economics that represents the maximum amount of money an individual is prepared to spend to acquire a specific good or service.

[9] https://datasetninja.com/isic-challenge-2018

[10] https://www.kaggle.com/datasets/altruistdelhite04/loan-prediction-problem-dataset

[11] https://myleott.com/op-spam.html

benign, accept, or genuine) and five for the negative class (i.e., malignant, reject, or deceptive). For each instance, we instantiated ten simulated AI systems with different fixed combinations of accuracy (low = 65%, or high = 90%) and data quality (low or high). We adjusted the accuracy values by adding a random noise with a range of 0-3% to the baseline values to simulate a refresh effect on the AI pool after each trial completion, further randomizing the order of the AI systems. Thus, we clearly indicate to the participants that the pool of AI systems is renewed after each trial. This design has several advantages. One, we ensure that subjects do not randomly identify a useful AI system at the beginning and then always stick with it. Second, we represent a large, dynamic market in which learning does not meaningfully alter the search set. Third, subjects learn about the market's lemon density through independent information signals across rounds. In our design, an AI that is a "lemon" delivers wrong suggestions 15% of the time, while a "peach" delivers correct suggestions 90% of the time. Consequently, the actual accuracy of the AI pool is bound to the lemon density: 70% for low, 40% for medium, and 10% for high density. To control for potential ordering biases [107], we (i) applied block randomisation across the presentation of the three tasks, and (ii) generated 400 random permutations of the ten instances to ensure each participant encountered uniquely ordered task instances.

3.6 Implementation and User Interface Design

An example of the interface participants interacted with during the study is shown in Figure 2. On each trial, users could hover over any AI to reveal the current state of information disclosure (see Fig. 2-A1). This was a deliberate design choice to validate participants' engagement with the available information on AI system quality in the disclosure conditions. Additionally, they could either delegate the decision to an AI (Fig. 2-B) or make the decision themselves (Fig. 2-C). After each decision, whether made with AI or independently, participants receive trial-by-trial feedback on the correctness of their choice, allowing them to adjust their strategies through the tasks. The online user interface was developed and hosted as a web application using the Next.js[12] framework and deployed on Vercel[13]. We used React[14] for the frontend implementation, tRCP[15] for the backend, and Supabase[16] for data storage with PostgreSQL.

4 Study Design

To answer our research questions, we conducted a pre-registered between-subjects study[17] comprising three *lemon density* conditions (**`Low Density`**, **`Medium Density`**, and **`High Density`**) × two *information disclosure* conditions (**`No Disclosure`** and **`Partial Disclosure`**), plus an additional condition with **`High Density`** and **`Full Disclosure`**, which we used as a benchmark. We do not gather complete factorial data for the **`Full Disclosure`** condition because, in theory, user behavior in this condition should be fully deterministic. Participants observe high-quality AI systems and always use them except if the participants believe themselves to be exceptionally good (i.e. ≥ 0.9 accuracy). Thus, **`Full Disclosure`** serves as a behavioral benchmark representing an "optimal" world in which information asymmetries can be institutionally solved. All data and code pertaining to the study are publicly available to ensure reproducibility.[18]

This section describes the variables and measurements collected during the study, as well as the recruitment policies for participants, statistical analysis setup, and study procedure.

4.1 Variables

4.1.1 *Independent Variables.*

- **Lemon density** (*categorical, between-subjects*). This variable controls the occurrences of AI "lemons" within the ten AIs pool for which participants can delegate the decisions to. We do not inform participants about the lemon density, although we do inform them that the share of AI systems remains constant throughout all tasks:
  - **`Low Density`**: 3 out of 10 AIs are lemons (2 low accuracy, 1 high accuracy) = 70% accuracy of AI pool.
  - **`Medium Density`**: 6 out of 10 AIs are lemons (4 low accuracy, 2 high accuracy) = 40% accuracy of AI pool.
  - **`High Density`**: 9 out of 10 AIs are lemons (6 low accuracy, 3 high accuracy) = 10% accuracy of AI pool.
- **Information disclosure** (*categorical, between-subjects*). This variable controls how much information is disclosed for each AI system, considering both simulated accuracy on the test set and the data quality on which an AI is trained:
  - **`No Disclosure`**: Participants only see AI names, as we do not disclose any information about each AI model.
  - **`Partial Disclosure`**: Participants only see the accuracy for each AI model.
  - **`Full Disclosure`**: Participants see the accuracy and data quality for each AI model.[19]

4.1.2 *Dependent Variables.*

- **Delegation to AI** (*continuous*). Percentage of participants' delegations to AI in the 30 trials.
- **Coins earned** (*continuous*). The number of coins participants earned as a result of correct predictions across the 30 trials (i.e., proxy of task performance).
- **Delegation to Lemon AI** (*continuous*). Percentage of participants' delegations to a "lemon AI" when they used the AI pool.

4.1.3 *Descriptive and Control Variables.*

- **Task familiarity** (*numerical, within-subjects*). We asked participants to state their familiarity with each task (i.e., loan prediction, deception detection, and skin cancer prediction) using a 5-point Likert scale from "1 - No experience" to "5 - Highly experienced".
- **Risk attitudes** (*numerical*). We assessed participants' risk attitudes using Dohmen et al.'s ten-point scale [41] by asking

[12] http://next.js/
[13] https://vercel.com/
[14] https://react.dev/
[15] https://trpc.io/
[16] Supabase.com
[17] The pre-registration can be found here: https://osf.io/95nbv/?view_only=031735e68ed04baeb54ba2c207fd7b13.
[18] Data and code for reproducibility: https://osf.io/qsaun/overview?view_only=42c81874313d4fb9899786fac847d24e
[19] Note that we exposed participants to the full disclosure condition only using high lemon density as a benchmark condition.

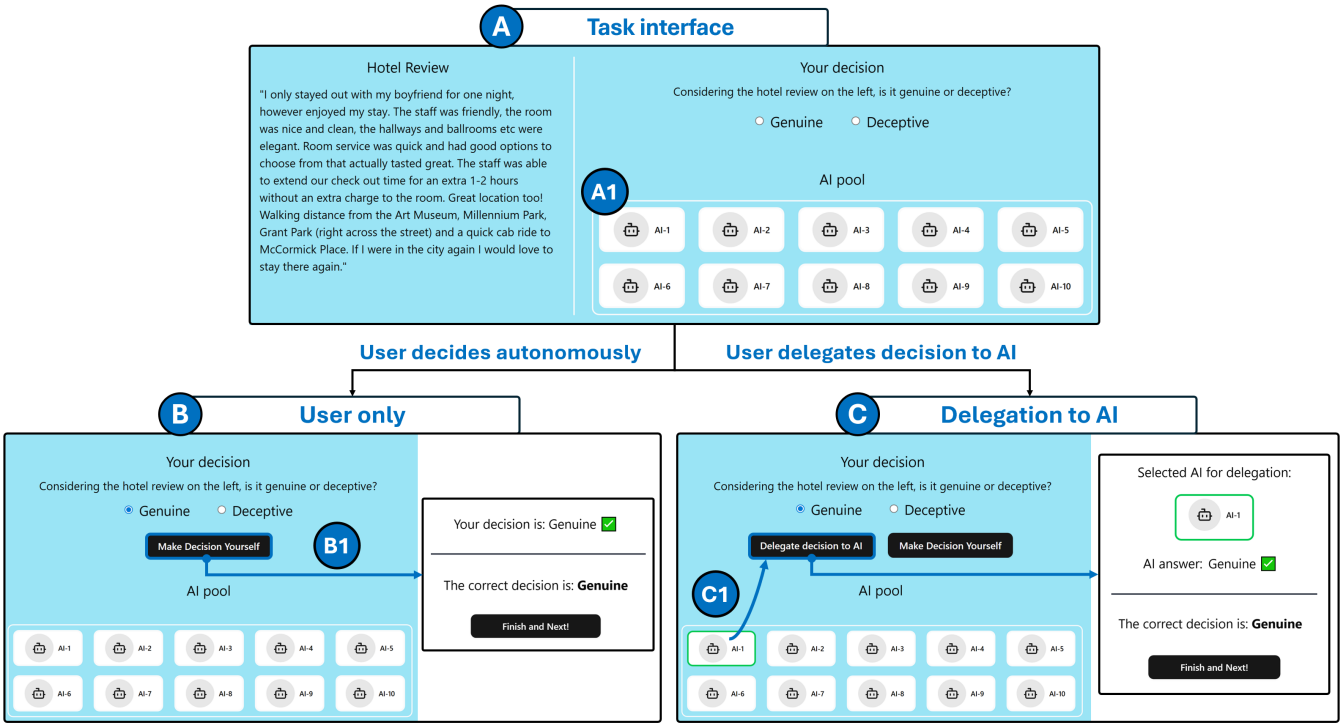

Figure 2: Participants' flow for each trial in our user study, using a deceptive review detection task and no disclosure condition (A), with a pool of AIs available at the bottom right of the interface (A1). Participants had two options for decision-making: they could either complete the decision themselves or delegate it to one of the AIs. In the "User only" option (B), users made their own decision and received correctness feedback by pressing the corresponding button (B1). Alternatively, in the "Delegation to AI" option (C), users hovered over the AIs in the pool to reveal their accuracy and data quality, if available, based on the experimental condition. Then, participants selected one AI (highlighted in green) and could press the corresponding button to delegate the decision to the AI (C1), thereby receiving its prediction and feedback about its correctness.

the following question: *"In general, how willing are you to take risks?"* (1: "not at all willing to take risks"; 10: "very willing to take risks.").

- **Affinity for technology** (*continuous*). This metric indicates the curiosity and willingness to engage with the technical workings of systems [86]. We measured it by taking the mean score of the four items (1: completely disagree, to 6: completely agree) in the Affinity for Technology Interaction (ATI) scale proposed by Franke et al. [52], consistent with prior work [86, 148, 149].
- **AI literacy** (*continuous*). Because people have different motivations to engage with AI assistance [37, 50, 89, 92, 149], we measured AI literacy by taking the mean of the four items that make up the self-assessed AI Literacy (AILIT) scale from Schoeffer et al. [125]. Responses were collected on a 5-point Likert scale (1: strongly disagree; 5: strongly agree).[20]
- **Perceived lemon density** (*numerical*). We asked participants to respond about how many AIs were lemons in their opinion on an eleven-point numerical scale (0-10) after the fourth and ninth trials of each task.

4.2 Participants

4.2.1 Recruiting and Filter Criteria. We received ethics approval from the German Association for Experimental Economics Research e.V. and then proceeded to recruit participants from Prolific[21] by following these selection criteria: equal participation across genders, age of 18 or more, high English proficiency, approval rate over 99%, and Desktop as a mandatory participation device. We rewarded participants with £2.5 for completing the study, based on an average completion time of 25 minutes, which equates to an average of £6 per hour. We gave an extra £0.1 to participants for each correct response to encourage high-quality work. Altogether, participants were compensated an average of £4.4 for a 25-minute completion time, which corresponds to an hourly wage of £10.56, which is considered a fair payment on the Prolific platform [79]. In total, we recruited 330 participants (50% female), 50 for each `No Disclosure` × Density and `Partial Disclosure` × Density conditions, as well as 30 for our benchmark `Full Disclosure` with `High Density` condition. We decided to recruit fewer participants in our benchmark condition because, theoretically, behavior in this

[20] Self-Assessed AI Literacy (AILIT) full questions: https://github.com/jakobschoeffer/facct22-130-appendix/blob/main/facct22-130-appendix.pdf

[21] https://www.prolific.com/

condition is fully deterministic. Herein, participants with an expected performance of < 0.9 would always choose a high-quality AI system and delegate.

4.3 Procedure

After obtaining informed consent, participants were assigned to one of the seven conditions based on the level of information disclosure and lemon density in a balanced fashion (except for the full disclosure, high lemon density condition, see Section 4.2). Next, we collected participants' familiarity with each of the three tasks, their risk attitudes, as well as their affinity for technology and AI literacy, as assessed by questionnaires. During this phase, they also completed two attention checks and could further proceed with the study only if they obtained at least one correct answer.[22] Then, they completed a familiarization tutorial that explained the general purpose of the study and important terminologies such as AI's accuracy and data quality. The tutorial included two trials for each task: skin cancer prediction, loan approval, and deceptive hotel reviews, delivered with block randomization ordering, which was also be used for the next 30 trials. The tutorial also outlined the option to choose an AI from a pool of ten to assist with current decisions, which could be accessed on demand by delegating the current decision to it. Therefore, after participants decide to select an AI for assistance during a trial, the pool of AI options will refresh for the next trial, while still respecting the current conditions related to information disclosure and lemon density (see Sec. 3.5).

Participants were then asked to answer some questions about the tutorial as comprehension checks, and only those who provided all the correct answers qualified to proceed with the user study. Afterwards, participants completed 30 trials, ten for each task, ensuring block randomization order across them. Specifically, upon entering a task, we provided participants with information about the dataset and the task's purpose, showing them one instance for the positive class and one for the negative class. For each decision, participants had the option to complete the trial themselves or delegate the trial's decision to an AI in the AI pool. Within each task, we additionally elicit subject beliefs about the market's lemon density after rounds four and nine. As a post-test, participants were informed about their total earned coins and asked for optional textual feedback about the study.

5 Results and Analysis

The final participants' sample comprised 330 users, with 166 females and 164 males, and an average age of $M = 37.44$ and $SD = 12.55$. Participants reported relatively low familiarity with all three tasks, and no differences were observed between the experimental conditions (see Table 5 in the Appendix). Similarly, participants' risk propensity did not differ across conditions (Kruskal-Wallis: $\chi^2 = 2.88$, $df = 6$, $p = 0.82$). Further, participants reported an average Affinity for technology of $M = 3.88$ ($SD = 0.58$, 6-point Likert scale) and an average AI literacy of $M = 3.81$ ($SD = 0.71$, 5-point Likert scale). [23]

First, we look at the effect of information asymmetries without disclosure institutions on user AI adoption and performance. Figure 3 shows average delegation shares and performance (Coins earned) across the seven conditions. In **`No Disclosure`**, delegation generally decreases with the density of lemons in the market. Over time (Figure 5 in the Appendix), we found that participants in **`Low Density`** increase delegation, but learning quickly stalls. In **`Medium Density`**, the trend is more ambiguous and even in **`High Density`**, the delegation shares remain comparably high. A mixed effect logistic panel regression (Table 6 in the Appendix) confirms only very limited treatment differences, where participants learn to delegate more (less) in **`Low Density`** (**`High Density`**), but effect sizes are small. This is surprising, because participants can learn over 30 rounds, and the differences between **`Low Density`** and **`High Density`** regarding expected value of AI delegation are quite large. Figure 6 confirms that participant beliefs λ systematically deviate from true density values and do not significantly improve over time. In **`Low Density`**, participants over-estimate the share of lemons, whereas they under-estimate them in **`Medium Density`** and **`High Density`**. Looking at **`Partial Disclosure`** in Figure 3 reveals that overall, delegation rates are similarly imprecisely tuned to the actual distribution of low-quality AI systems. Despite similar delegation rates, the right panel shows that participants in **`Partial Disclosure`** substantially outperform those in **`No Disclosure`** with regard to performance, providing support for hypothesis H4. To quantify the effect of partial disclosure institutions on participants' delegation behavior, we run a mixed effects logistic regression with participant-level random effects (Table 2), confirming no effect of **`Partial Disclosure`**. However, in line with Bayesian information updating, partial disclosure does lead to significant and large efficiency increases (Table 3) in **`Low Density`** and **`Medium Density`**. Thus, disclosing accuracy information does not affect the prevalence of delegation, but the quality. Finally, in our benchmark condition **`Full Disclosure High Density`**, participants only delegate 57.7% of problems to the AI, despite being able to identify a high-quality system with certainty. This leads to substantial efficiency losses of 147 Coins on average.

> **Result 1:** Under full information asymmetries, participants exhibited limited learning about the prevalence of low-quality AI systems. This leads to user adjustments in AI adoption that are directionally correct, but substantially too small.
>
> **Result 2:** Partial disclosure institutions increase the efficiency of Human-AI collaboration under information asymmetries as long as the density of lemons is not too high.

To test the proposed mechanism of our model, Figure 4 illustrates the share of delegation choices in which participants selected a lemon AI. For **`No Disclosure`**, the percentage is close to the actual distribution of lemons. Participants in **`Partial Disclosure`**, on the other hand, are much more likely to avoid the low-quality systems across *all* disclosure conditions. A random effects panel logit regression (Table 4) confirms that the partially informative

[22] We used valid attention checks allowed from Prolific where the answer was explicitly reported in the question text: (i) "For this question, please select the option "Largely disagree""; (ii) "For this question, please select the option "Largely agree"".

[23] Shapiro-Wilk tests indicated significant deviation from normality for both measures (AI literacy: $W = 0.966$, $p < .0001$; affinity for technology: $W = 0.954$, $p < .0001$). Kruskal-Wallis tests showed no significant differences across conditions for AI literacy ($\chi^2(6) = 2.74$, $p = .84$), whereas affinity for technology differed across conditions ($\chi^2(6) = 13.4$, $p = .037$). Post-hoc pairwise comparisons (Bonferroni-adjusted) identified a single significant contrast for affinity for technology: *full_high* vs *none_medium* ($Z = 3.10$, $p = .0019$, $p_{\text{adj}} = .0403$).

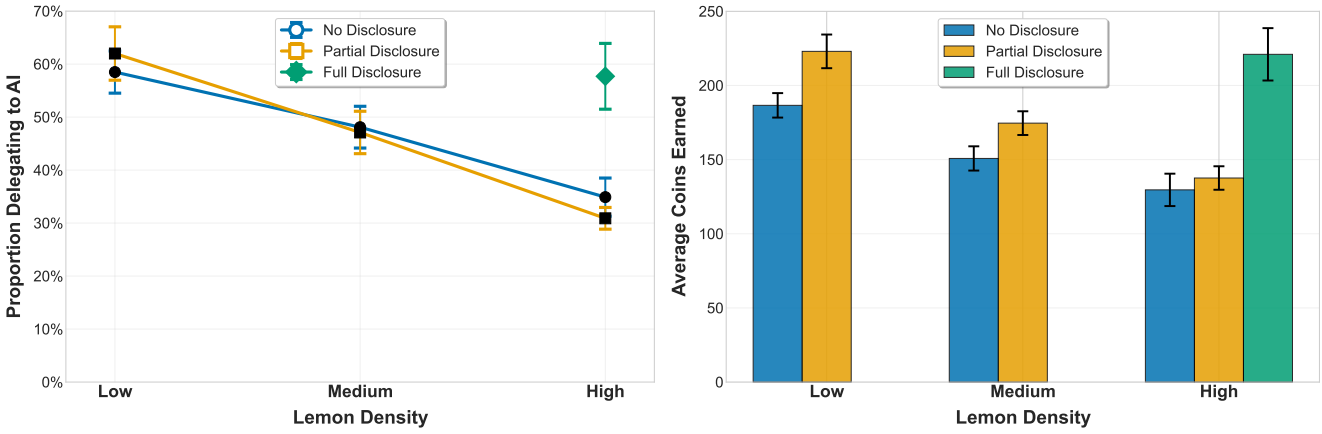

Figure 3: Left: Average delegation rates across the seven conditions collapsed at task-level. There are no significant differences in delegation to AI between the `No Disclosure` and `Partial Disclosure` conditions across the three lemon densities, whilst `Full Disclosure` only outcompetes `No Disclosure` in terms of delegation rate in the *High density* condition. Right: Average coins earned per task across the seven conditions. Participants in `Partial Disclosure` outperform those in `No Disclosure` in terms of the number of coins earned, except for the `High Density` condition. Instead, the average coins earned in the `Full Disclosure` condition is significantly higher than `No Disclosure` and `Partial Disclosure` conditions with a `High Density` of lemons. Error bars represent 95% confidence intervals.

Table 2: Mixed-effects logistic regression of delegation to AI (log-odds). Random intercept by participant. Task and time fixed effects included. Baseline disclosure is *No Disclosure*.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	0.580 (0.495)	-0.113 (0.343)	-0.123 (0.222)	-0.103 (0.291)
Full	—	—	—	1.288 (0.343)***
Task FE (domain)	Yes	Yes	Yes	Yes
Time FE (round)	Yes	Yes	Yes	Yes
Observations	3,000	3,000	3,000	3,900

Notes. Coefficients are log-odds; standard errors in parentheses. Models are GLMMs with logit link, random intercept for participant (1|user_id), task fixed effects (domain dummies), and time fixed effects (round dummies). Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

accuracy labels allow participants to be significantly more likely to avoid AI lemons. People are generally able to improve their decision-making even when they only receive incomplete disclosure information. However, this effect appears to decrease in the density of low-quality AIs. While lemon shares are basically optimal in **`Low Density`**, they are 15 percentage points "too high" in **`Medium Density`** and 55 percentage points "too high" in **`High Density`**. This could be due to selection effects, where more engaged or capable participants quickly learn to never delegate in **`High Density`**, or cognitive effort, which plausibly increases with the share of ambiguous accuracy label signals and thus potentially inhibits the efficiency of the Bayesian updating process in the context of information asymmetries.

> **Result 3:** In line with the predictions of Bayesian updating under information asymmetries, participants who receive partial disclosure information are significantly less likely to rely on a lemon AI.

> **Result 4:** The positive effect of partial information disclosure on AI system selection decreases with the density of low-quality AI systems.

Now, we focus on the rationality of participant choices, assuming risk neutrality. We are primarily interested in whether partial disclosure causes diverging delegation trends between participants who outperform and those who under-perform the AI market (i.e. expected-payoff maximizing behavior), or if performance improvements are due to more targeted delegation choices independent from the user's accuracy. Because participants may have different abilities across tasks, we run task-separate mixed-effect logit panel regressions. Our main target variable interacts the disclosure condition with a dummy variable capturing whether a participant's expected payoff based on their own performance exceeds the expected payoff of entering the market (and making an informed choice in **`Partial Disclosure`** and **`Full Disclosure`**). All tables are shown in the Appendix (Tables 7 – 9). The interaction term does

Table 3: OLS (user-clustered) regressions of task coins by disclosure and density. Task fixed effects included; baseline disclosure is *No Disclosure*.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	36.400*** (7.159)	23.800*** (5.835)	8.000 (6.878)	8.000 (6.871)
Full	—	—	—	91.400*** (10.531)
Task FE (domain)	Yes	Yes	Yes	Yes
SE type	Clustered (user)	Clustered (user)	Clustered (user)	Clustered (user)
Observations	300	300	300	390

Notes. Coefficients; cluster-robust standard errors in parentheses. Baseline disclosure is *No*. Models include task fixed effects (domain dummies). Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 4: Mixed-effects logistic regression of selecting a lemon among delegations (log-odds). Random intercept by participant. Task and time fixed effects included. Baseline disclosure is *No Disclosure*.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	−2.377*** (0.235)	−0.926*** (0.115)	−0.803*** (0.222)	−0.868* (0.362)
Full	—	—	—	−5.807*** (0.464)
Task FE (domain)	Yes	Yes	Yes	Yes
Time FE (round)	Yes	Yes	Yes	Yes
Observations	1,808	1,428	988	1,507

Notes. Coefficients are log-odds; standard errors in parentheses. Models: GLMM (logit) with random intercept for participant (1|user_id), task fixed effects (domain), and time fixed effects (round). Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

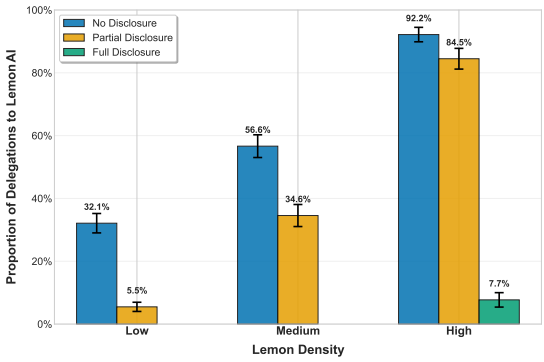

Figure 4: Share of delegation choices that targeted a lemon AI system. The bars only include choice data from observations in which a participant used the market (error bars represent 95% confidence intervals). Participants in the `Partial Disclosure` condition delegated significantly fewer decisions to lemon AIs than in the `No Disclosure` condition across all lemon density conditions. Participants in the `Full Disclosure`condition outperformed those in both `No Disclosure`/`Partial Disclosure` in terms of delegation rate in the `High Density` condition.

not have explanatory value for any of the three tasks, suggesting that partial disclosure information does not increase the rationality of users. Instead and in line with the delegation behavior plotted in Figure 3, both the delegation shares and the delegation distribution across skill levels is largely equivalent between **`No Disclosure`** and **`Partial Disclosure`**. Hence, performance improvements are driven by *more targeted* delegation choices, where participants use the accuracy label to successfully avoid lemon products. Beyond that, Tables 7 – 9 also confirm that participants with more accurate predictions than the AI market are generally less likely to use an AI system for deceptive hotel reviews and loan prediction. Thus, participants are sensitive to their own relative performance level during delegation choices. Regarding "blind" utilization of the market where participants always delegate, an OLS regression with clustered standard errors on the participant level (Table 10) finds very limited evidence for treatment differences. Mostly and in line with expectations, **`Full Disclosure`** leads to a significant increase of full delegators. This makes sense, because participants can perfectly identify a high-quality model. **`Partial Disclosure`** shows no consistent effect, but may increase full market usage in **`Low Density`**, i.e. when the density of high-quality AIs is high.

Result 5: Partial disclosure institutions do not affect the share of participants who maximize their expected payoff.

Result 6: Partial disclosure institutions increase average user performance and market efficiency by allowing delegators to make more informed delegation choices.

Finally, we analyze how information asymmetries and participant-level characteristics affect AI use and Human-AI collaboration efficiency more generally. To that end, we deploy a full mixed effect logistic panel regression with all conditions and our main co-variates technology affinity, risk attitudes, AI literacy, and task familiarity. Results (Table 11) confirm that delegation decreases with lemon density, thus supporting hypothesis H2, but increases over the course of each task. Participants adapt directionally correct, albeit too little, and learn over time. In addition, we find significant negative correlations between task familiarity and delegation, as well as affinity for technology and delegation, while risk attitudes and AI literacy do not explain behavior. Looking at performance (Table 12), in contrast, suggests that participants with more risk-loving preferences tend to make fewer correct predictions, while AI literacy correlates positively with income. Panel 2 adds performance and delegation variables, which illustrate that people tend to over-rely on AI in **`No Disclosure`** and thereby reduce their potential income. For **`Partial Disclosure`**, on the other hand, participants under-rely on AI in **`Low Density`**, and over-rely in **`High Density`**. Notably, increasing the lemon density to medium does not negatively affect user performance under the partial information disclosure condition, underscoring the usefulness of the partial signal, as it almost entirely offsets the increase in low-quality systems. With high density, people always strongly over-delegate in the absence of full disclosure. Overall, participant choices are only imperfectly calibrated to the decision environment under information asymmetries. They delegate too little if lemons are rare, and delegate too much when lemons are common. Participants efficiently exploit partial information signals, causing more targeted delegation choices. However, rationality does not improve. Fully disclosing all relevant variables has a strong and positive effect on efficiency, but is severely constrained by participants' tendency to rely on themselves.

6 Discussion

This paper experimentally analyzes the effect of information asymmetries in the presence of low-quality AI systems (*lemons*), on user behavior, and the concurrent efficacy of (imperfect) disclosure. It is notoriously difficult to solve informational market frictions through regulatory interventions, as full transparency is often neither an achievable, nor desirable goal. The market for AI systems illustrates this problem very well. Regulations differ substantially between countries, leave a lot of room for interpretation, and are subject to constant debates between politicians, regulators, providers and citizens. In addition, there has so far been very little empirical work about the behavioral validity of different interventions, such as (partial) disclosure rules (see, e.g., EU AI Act), which adds to the general confusion of the debate. Specifically in the context of disclosure, one main driver behind the efficacy of forcing or encouraging providers to reveal relevant information is the behavior of users. In order to be useful, people must be able to integrate the relevant information signals into their choice process. This is, arguably, a lot easier when users have full information about all AI systems, because it enables simple cost-benefit comparisons. In reality, however, people must handle a lot more ambiguity, and follow imprecise signals that may affect optimal choices, but are not sufficiently complete to render behavior deterministic. Hence, improvements are subject to users' belief updating processes and, by extension, rationality. This tension is the main focus of our paper.

RQ1: How do information asymmetries about AI system capabilities affect users' adoption of AI and market outcomes in the presence of low-quality AI systems?

Information asymmetries are detrimental to the outcomes and the overall efficiency of Human-AI collaboration. Participants in our study over-reacted to uncertainty when the density of lemons was low, and under-reacted when the density was high. The quality of delegation choices improved at the beginning, but quickly stalled, suggesting limited learning under uncertainty. In line with this, participants' beliefs about the market's lemon density were stable over time (Figure 6). Under- and over-utilization of the AI systems was common. As a result, utilization of low-quality AI systems remained high, and efficiency is mostly a function of lemon density. A central finding of this work is the persistent under-reliance on AI systems even in the **`Full Disclosure`** condition, where participants could identify a high-quality AI system with certainty. This suggests that transparency alone is insufficient to guarantee optimal adoption. Several psychological and cognitive factors may explain this behavior. First, overconfidence in self-judgment can lead users to favor their own reasoning over algorithmic recommendations [64], even when objective evidence indicates superior system performance. Second, users may exhibit a strong preference for agency and control, consistent with prior HCI research showing that autonomy is often prioritized over efficiency in decision-making contexts [98]. Third, residual distrust of AI systems, stemming from concerns about hidden risks or accountability, may persist even when technical information is fully disclosed, reflecting challenges in trust calibration.

These explanations align with established frameworks such as trust in automation [88, 98], which emphasize that trust is shaped by more than objective reliability, and bounded rationality [78], which highlights cognitive effort and satisficing strategies under uncertainty [124]. From an HCI perspective, this finding underscores the need for interaction designs that go beyond transparency to actively support trust calibration and reduce cognitive burden.

RQ2: How do different information disclosure requirements about AI system capabilities impact user behavior, market outcomes, and reliance on low-quality AI systems?

We document four main results. One, partial disclosure substantially increases the efficiency of human-AI collaboration in the presence of low-quality AI systems. Two, this effect is driven by more effective delegation through the marginalization of low-quality AI systems, rather than increased rationality. Over- and under-delegation rates remain constant, but delegation overall improves. Three, high density of low-quality AI systems alleviates the positive effect of partial disclosure, as participants are significantly less effective in utilizing the provided information. Fourth, even under

a full disclosure institution that entirely eliminates market information asymmetries, participants exhibit strong AI under-reliance, leading to large efficiency losses in decision-making.

AI qualities as a vehicle for information disclosure. In this work, we leverage AI accuracy as a cue for partial disclosure and incorporate additional data quality indicators to enable full disclosure. These choices reflect the need for interpretable, actionable signals that help lay users make informed decisions during interactions. Our approach is generalizable and grounded in human-centered AI literature on decision-making, where model-centric cues (e.g., AI accuracy, confidence, and calibration) are widely recognized for their influence on user reliance attitudes [28, 30, 63, 76, 118, 147, 151]. In contrast, "data-centric explanations", which foreground properties of the training data, have only recently emerged in XAI literature. These explanations, when combined with model-centric ones, can form hybrid strategies that support richer, more transparent interactions, helping foster appropriate reliance on AI systems [7, 15, 16, 30, 150]. Future work should investigate how different combinations of partial- and full-disclosure cues shape user perceptions, trust calibration, and delegation behaviors across diverse contexts, tasks, and stakes. Such studies will be critical for designing disclosure mechanisms that reduce information asymmetry and support equitable, user-centered decision-making in real-world interactive systems.

6.1 Implications

Our work bears important implications for the future design and regulation of AI systems. As AI systems proliferate across consumer and professional contexts, our results underscore the critical role of information disclosure in shaping user reliance behaviors. Our insights demonstrate that information asymmetries between system providers and end-users do not merely create market inefficiencies, but they also lead to miscalibrated trust, poor delegation strategies, and suboptimal task outcomes. Users in our study rarely learned enough to identify environments with high versus low prevalence of low-quality systems, resulting in persistent patterns of over-reliance or under-reliance. This corroborates recent work in the HCI community that has argued that misaligned mental models can undermine trust and decision-making, and that grounding interactions in shared understanding is essential [5, 6, 11, 65].

Our findings suggest that even partial disclosures can meaningfully improve user decision-making, provided they are interpretable and actionable. HCI researchers and practitioners must prioritize mechanisms that surface AI accuracy, uncertainty, provenance, and limitations, even if the information eventually disclosed to users of such systems is incomplete. From a regulatory and policy perspective, prioritizing enforceable disclosure rules, even if incomplete, can be important. Policy and regulatory efforts could aim to enforce minimum disclosure standards for AI systems that mandate interpretable cues (e.g., accuracy and data provenance) rather than exhaustive technical details, which can also be difficult to enforce or audit. Preventing selective reporting or "gaming" of transparency by AI system developers, providers, or suppliers would require enforceable strategy-proof formats (e.g., standardized templates, or machine-readable labels).

From the perspective of design implications for human-AI interaction, our findings underline the importance of embedding meaningful information disclosure cues near decision points (e.g., alongside AI advice or assistance) rather than in separate documentation or other pathways. For example, providing short tooltips or expandable panels that explain to users what accuracy and data quality mean for the task at hand can help surface the required disclosure signals in situ. Filtering mechanisms can enable users to select AI systems based on quality tiers, supporting user agency and control. In addition, incorporating dynamic feedback mechanisms that respond to user behavior (e.g., nudging users when persistent under-reliance is detected, or highlighting missed opportunities for efficiency gains) can help support trust calibration without forcing delegation.

Collectively, these results reinforce the HCI imperative to design for informed interaction [9, 31, 129, 134], where users can meaningfully interpret and act upon AI outputs, even in the presence of systemic opacity. Rather than aiming for exhaustive transparency, designers and regulators should prioritize strategically minimal disclosures that help users form accurate mental models.

6.2 Caveats, Limitations, and Future Work

Static Market Dynamics. In this study, we rely on a market without dynamic exit and entrance of sellers and consumers, as well as fixed seller behavior. By holding market dynamics constant, we can focus on how users interpret and act on the disclosure signals in a controlled environment. This provides actionable insights for designing trust cues, information labels, and decision-support interfaces that help users make informed choices under uncertainty. However, it is worth noting that several real-world AI marketplaces are dynamic, in which users' choices can influence sellers' behavior and vice versa. While our setup allows us to focus on consumer behavior, long-term dynamics, price-setting and market equilibria are beyond the scope of this paper. Future research may expand on our results and framework to gather more nuanced and generalizable results through market experiments. Future work should also explore interactive and adaptive disclosure designs that remain effective when market conditions change, ensuring transparency and appropriate adoption of AI systems at scale.

Operationalization of AI Quality. Our study relies on simulated AIs with relatively fixed accuracies with only two relevant quality dimensions (accuracy on the test set and data quality in the training set). While performance cues and training data quality are powerful and simple signals for lay users when considering AI models (e.g., similar to model cards or AI nutrition labels [13, 55, 80, 103, 128, 136]), future work may expand on our results by deploying "real" AI systems and introducing higher-dimensional quality attributes. This, in particular, may interact with how people process information signals, as they directly affect the complexity and kind of information users observe. However, increasing AI quality cues should account for both lay and expert users. Presenting too many AI qualities may overwhelm lay users and increase overconfidence biases such as inflated self-assessment (e.g., the Dunning-Kruger effect [64]), leading to misunderstandings and inappropriate reliance on AI. Expert users, instead, can benefit from

richer information, including data inspection (e.g., intrinsic data biases, data quality, and processing) [7, 16, 30] and AI indicators (e.g., bias propagated to the model, fairness principles, and confidence calibration) [28, 95, 130, 144]. In addition, further studies should test AI quality disclosures across different task stakes and difficulties to understand how various indicators influence real-world delegation and market outcomes.

On the nature of 'lemons.' While our study operationalizes low-quality AI systems (i.e., lemons) through accuracy and data quality, real-world AI systems exhibit additional quality dimensions such as fairness, robustness, and safety. These characteristics are often harder to quantify and communicate because they depend on context, edge cases, or adversarial conditions. If lemons were defined by these properties, the dynamics of user interaction with the AI systems and their reliance on AI advice might shift significantly. Users could underestimate risks even under full disclosure, or struggle to interpret complex ethical trade-offs. This would imply that disclosure mechanisms must evolve beyond numeric performance indicators to include interpretable signals for fairness and safety; potentially through scenario-based explanations, visual risk cues, or tiered disclosure interfaces. This is an important avenue for future research. Broadening the scope of quality dimensions is essential for designing transparency strategies that support informed human-AI interaction in ethically sensitive domains.

Task complexity. In our study, the decision-making tasks were cognitively manageable, which may have allowed participants to devote more attention to evaluating the information disclosed about the AI systems. In more complex or constrained contexts (e.g., under time pressure) [27, 122, 123, 137], users may have fewer cognitive resources available for scrutinizing AI quality signals, which could amplify the negative effects of information asymmetry and even diminish the benefits of full disclosure.

Delegation. We made a deliberate choice to model delegation as a binary decision (i.e., relying on AI or not). This offers a clear, interpretable starting point for studying reliance behaviors under controlled conditions. This simplification is common in human-centered AI research [20, 47] because it allows us to isolate the core phenomenon (information asymmetry and disclosure effects) without introducing confounding factors from more complex interaction patterns. We can thereby systematically test how disclosure strategies and lemon density influence adoption decisions, which would be difficult to disentangle in a continuous reliance model. Moreover, binary delegation reflects real-world decision points in many contexts, such as whether to accept an AI recommendation or proceed manually (e.g., approving a credit score, accepting a medical triage suggestion). While actual workflows often include verification or partial reliance, these behaviors are layered on top of the fundamental choice to delegate or not. Understanding this baseline is essential before designing for more nuanced behaviors. Having said that, future research should explore experimental designs that allow for graded reliance (e.g., sliders for confidence weighting), verification actions (e.g., optional checks before committing), and override mechanisms to represent other real-world workflows.

Furthermore, additional metrics focusing on different users' perceptions and traits could provide a more nuanced understanding of our results. For example, in light of participants' under-delegation in the full disclosure high-density condition, future research could incorporate measures of individual characteristic that are relevant for explaining delegation behavior, such as human decision confidence [30, 51, 90, 101, 111], perceived trust and autonomy [61, 94, 95], or algorithm aversion (i.e., the tendency to reject the suggestions made by algorithms, even when those algorithms outperform human judgment) [40, 56, 75]. This, along with people's actual ability to appropriately rely on AI systems, may induce profound changes in the effectiveness of disclosure regimes.

Costless Verification. Real world disclosure regimes are noisy, costly, and exhibit variance in credibility and trust. We abstract from these factors, but acknowledge that these factors will play a role for actual disclosure regulations. This includes sellers' strategic verification choices, price-setting, endogenous seller competition via consumer choices, and, in equilibrium, what kind of disclosure level emerges from strategic market interactions. It is, for instance, not guaranteed that in equilibrium, people can fully trust partially informative information labels (in our case, we select a publicly known and fixed probability of a wrong signal, which reveals the usefulness of the label fully).

Future research should investigate why users under-delegate even under full disclosure. Qualitative or mixed-methods studies could unpack underlying factors such as trust calibration, perceived control, and accountability concerns, offering richer insights into cognitive and social dynamics that shape reliance on AI [54]. Building on these findings, we aim to translate behavioral patterns into actionable interaction designs in the near future. For example, tiered disclosure interfaces that progressively reveal information [133], or adaptive trust cues that respond to user behavior in real time. Expanding our operationalization of AI quality beyond accuracy and data generalizability to include fairness, robustness, and safety will further align disclosure strategies with emerging ethical and technical standards [45, 87].

7 Conclusion

This paper experimentally demonstrates that information asymmetries fundamentally undermine efficient AI adoption in consumer markets. When faced with uncertainty about AI quality across a pool of opaque systems, participants adjust reliance in the right direction over time, but still under-use AI when lemons are rare and over-use AI when lemons are common. Partial disclosure markedly improves efficiency by steering delegators away from lemons, even though it does not increase the share of "rational" adopters. Under full disclosure, users still under-delegate, leaving sizable performance on the table. Together, these findings suggest that enforceable, even imperfect, disclosure rules can deliver meaningful welfare gains in real-world AI markets that suffer from information asymmetries. They underscore the critical role of institutional design in shaping AI adoption patterns and highlight the potential for targeted regulatory interventions to mitigate market failures in the growing AI economy. Our work introduces a novel experimental framework that adapts the classic "market for lemons" theory to study AI adoption under information asymmetry. By operationalizing lemon density as a design variable, we systematically examine how disclosure strategies interact with uncertainty to shape user reliance. This approach not only bridges economic theory and HCI

but also offers an extensible paradigm for future research on trust calibration, transparency, and the design of human interaction with AI systems.

Acknowledgments

We thank all the anonymous participants in our study. This work was partially supported by the TU Delft AI Initiative, the Model Driven Decisions Lab (*MoDDL*), the Robust LTP GENIUS Lab, and the *ProtectMe* Convergence Flagship.

References

[1] Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz. 2023. *Combining human expertise with artificial intelligence: Experimental evidence from radiology.* Technical Report. National Bureau of Economic Research.

[2] George A Akerlof. 1970. The market for "lemons": quality uncertainty and the market mechanism. *The Quarterly Journal of Economics* 84, 3 (1970), 488–500.

[3] Mir Masood Ali, David G Balash, Monica Kodwani, Chris Kanich, and Adam J Aviv. 2023. Honesty is the Best Policy: On the Accuracy of Apple Privacy Labels Compared to Apps' Privacy Policies. *arXiv preprint arXiv:2306.17063* (2023).

[4] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. 2023. Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion* 99 (2023), 101805. doi:10.1016/j.inffus.2023.101805

[5] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.

[6] Robert W Andrews, J Mason Lilly, Divya Srivastava, and Karen M Feigh. 2023. The role of shared mental models in human-AI teams: a theoretical review. *Theoretical Issues in Ergonomics Science* 24, 2 (2023), 129–175.

[7] Ariful Islam Anik and Andrea Bunt. 2021. Data-Centric Explanations: Explaining Training Data of Machine Learning Systems to Promote Transparency. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 75, 13 pages. doi:10.1145/3411764.3445736

[8] Siddhant Arora, Danish Pruthi, Norman Sadeh, William Cohen, Zachary Lipton, and Graham Neubig. 2022. Explain, Edit, and Understand: Rethinking User Study Design for Evaluating Model Explanations. *Proceedings of the AAAI Conference on Artificial Intelligence* 36 (06 2022), 5277–5285. doi:10.1609/aaai.v36i5.20464

[9] Tita Alissa Bach, Amna Khan, Harry Hallock, Gabriela Beltrão, and Sonia Sousa. 2024. A systematic literature review of user trust in AI-enabled systems: An HCI perspective. *International Journal of Human–Computer Interaction* 40, 5 (2024), 1251–1266.

[10] Nagadivya Balasubramaniam, Marjo Kauppinen, Antti Rannisto, Kari Hiekkanen, and Sari Kujala. 2023. Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology* 159 (2023), 107197.

[11] Gagan Bansal, Besmira Nushi, Ece Kamar, Walter S Lasecki, Daniel S Weld, and Eric Horvitz. 2019. Beyond accuracy: The role of mental models in human-AI team performance. In *Proceedings of the AAAI conference on human computation and crowdsourcing*, Vol. 7. 2–11.

[12] Catarina Barata, Veronica Rotemberg, Noel C. F. Codella, Philipp Tschandl, Christoph Rinner, Bengu Nisa Akay, Zoe Apalla, Giuseppe Argenziano, Allan Halpern, Aimilios Lallas, Caterina Longo, Josep Malvehy, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2023. A reinforcement learning model for AI-based decision support in skin cancer. *Nature Medicine* 29 (2023), 1941–1946. doi:10.1038/s41591-023-02475-5

[13] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. doi:10.1162/tacl_a_00041

[14] Astrid Bertrand, Rafik Belloum, James R Eagan, and Winston Maxwell. 2022. How cognitive biases affect XAI-assisted decision-making: A systematic review. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 78–91.

[15] Aditya Bhattacharya, Jeroen Ooge, Gregor Stiglic, and Katrien Verbert. 2023. Directive Explanations for Monitoring the Risk of Diabetes Onset: Introducing Directive Data-Centric Explanations and Combinations to Support What-If Explorations. In *Proceedings of the 28th International Conference on Intelligent User Interfaces* (Sydney, NSW, Australia) *(IUI '23)*. Association for Computing Machinery, New York, NY, USA, 204–219. doi:10.1145/3581641.3584075

[16] Aditya Bhattacharya, Simone Stumpf, Lucija Gosak, Gregor Stiglic, and Katrien Verbert. 2024. EXMOS: Explanatory Model Steering through Multifaceted Explanations and Data Configurations. In *Proceedings of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 314, 27 pages. doi:10.1145/3613904.3642106

[17] Aditya Bhattacharya, Simone Stumpf, and Katrien Verbert. 2024. An Explanatory Model Steering System for Collaboration between Domain Experts and AI. In *Adjunct Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP Adjunct '24)*. Association for Computing Machinery, New York, NY, USA, 75–79. doi:10.1145/3631700.3664886

[18] Jan Biermann, John J Horton, and Johannes Walter. 2022. Algorithmic advice as a credence good. *ZEW-Centre for European Economic Research Discussion Paper* 22-071 (2022).

[19] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3173951

[20] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2025. Mind the Gap! Choice Independence in Using Multilingual LLMs for Persuasive Co-Writing Tasks in Different Languages. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025-1 May 2025*, Naomi Yamashita, Vanessa Evers, Koji Yatani, Sharon Xianghua Ding, Bongshin Lee, Marshini Chetty, and Phoebe O. Toups Dugas (Eds.). ACM, 937:1–937:20. doi:10.1145/3706598.3713201

[21] Shreyan Biswas, Alexander Erlei, and Ujwal Gadiraju. 2026. Belief Updating and Delegation in Multi-Task Human–AI Interaction: Evidence from Controlled Simulations. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*.

[22] Edyta Bogucka, Marios Constantinides, Sanja Šćepanović, and Daniele Quercia. 2024. AI Design: A Responsible AI Framework for Impact Assessment Reports. *IEEE Internet Computing* (2024).

[23] Bryan Bollinger, Phillip Leslie, and Alan Sorensen. 2011. Calorie Posting in Chain Restaurants. *American Economic Journal: Economic Policy* 3, 1 (2011), 91–128. doi:10.1257/pol.3.1.91

[24] Clara Bove, Jonathan Aigrain, Marie-Jeanne Lesot, Charles Tijus, and Marcin Detyniecki. 2022. Contextualization and Exploration of Local Feature Importance Explanations to Improve Understanding and Satisfaction of Non-Expert Users. In *27th International Conference on Intelligent User Interfaces* (Helsinki, Finland) *(IUI '22)*. Association for Computing Machinery, New York, NY, USA, 807–819. doi:10.1145/3490099.3511139

[25] Catherine Burrell. 2016. How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3, 1 (2016), 2053951715622512. doi:10.1177/2053951715622512

[26] Ángel Alexander Cabrera, Adam Perer, and Jason I. Hong. 2023. Improving Human-AI Collaboration With Descriptions of AI Behavior. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW1, Article 136 (April 2023), 21 pages. doi:10.1145/3579612

[27] Shiye Cao, Catalina Gomez, and Chien-Ming Huang. 2023. How time pressure in different phases of Decision-Making influences Human-AI collaboration. *Proceedings of the ACM on Human-computer Interaction* 7, CSCW2 (2023), 1–26.

[28] Shiye Cao, Anqi Liu, and Chien-Ming Huang. 2024. Designing for Appropriate Reliance: The Roles of AI Uncertainty Presentation, Initial User Decision, and User Demographics in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW1, Article 41 (April 2024), 32 pages. doi:10.1145/3637318

[29] Trent N Cash, Daniel M Oppenheimer, Sara Christie, and Mira Devgan. 2025. Quantifying uncert-AI-nty: Testing the accuracy of LLMs' confidence judgments. *Memory & Cognition* (2025), 1–26.

[30] Federico Maria Cau and Lucio Davide Spano. 2025. The Influence of Curiosity Traits and On-Demand Explanations in AI-Assisted Decision-Making. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 1440–1457. doi:10.1145/3708359.3712165

[31] Luciano Cavalcante Siebert, Maria Luce Lupetti, Evgeni Aizenberg, Niek Beckers, Arkady Zgonnikov, Herman Veluwenkamp, David Abbink, Elisa Giaccardi, Geert-Jan Houben, Catholijn M Jonker, et al. 2023. Meaningful human control: actionable properties for AI system development. *AI and Ethics* 3, 1 (2023), 241–255.

[32] Tirtha Chanda, Katja Hauser, Tabea-Clara Bucher, Carina Nogueira Garcia, Christoph Wies, Eva Krieghoff-Henning, Titus J. Brinker, Reader Study Consortium, et al. 2024. Dermatologist-like explainable AI enhances trust and confidence in diagnosing melanoma. *Nature Communications* 15 (2024), 524. doi:10.1038/s41467-023-43095-4

[33] Valerie Chen, Q. Vera Liao, Jennifer Wortman Vaughan, and Gagan Bansal. 2023. Understanding the Role of Human Intuition on Reliance in Human-AI Decision-Making with Explanations. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 370 (oct 2023), 32 pages. doi:10.1145/3610219

[34] Zhihan Cheng, Yue Wu, Yule Li, Lingfeng Cai, and Baha Ihnaini. 2025. A Comprehensive Review of Explainable Artificial Intelligence (XAI) in Computer Vision. *Sensors* 25, 13 (2025), 4166. doi:10.3390/s25134166

[35] Ben Chester Cheong. 2024. Transparency and accountability in AI systems: safeguarding wellbeing in the age of algorithmic decision-making. *Frontiers in Human Dynamics* 6 (2024), 1421273.

[36] Raj Chetty, Adam Looney, and Kory Kroft. 2009. Salience and Taxation: Theory and Evidence. *American Economic Review* 99, 4 (2009), 1145–1177. doi:10.1257/aer.99.4.1145

[37] Chun-Wei Chiang and Ming Yin. 2021. You'd Better Stop! Understanding Human Reliance on Machine Learning Models under Covariate Shift. In *Proceedings of the 13th ACM Web Science Conference 2021* (Virtual Event, United Kingdom) *(WebSci '21)*. Association for Computing Machinery, New York, NY, USA, 120–129. doi:10.1145/3447535.3462487

[38] Avishek Choudhury and Zaira Chaudhry. 2024. Large language models and user trust: consequence of self-referential learning loop and the deskilling of health care professionals. *Journal of Medical Internet Research* 26 (2024), e56764.

[39] Michael Chromik, Malin Eiband, Felicitas Buchner, Adrian Krüger, and Andreas Butz. 2021. I Think I Get Your Point, AI! The Illusion of Explanatory Depth in Explainable AI. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 307–317. doi:10.1145/3397481.3450644

[40] Berkeley Dietvorst, Joseph Simmons, and Cade Massey. 2014. Algorithm Aversion: People Erroneously Avoid Algorithms After Seeing Them Err. *Journal of Experimental Psychology: General* 144 (11 2014), 114–126. doi:10.1037/xge0000033

[41] Thomas Dohmen, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences. *Journal of the European Economic Association* 9, 3 (06 2011), 522–550. doi:10.1111/j.1542-4774.2011.01015.x arXiv:https://academic.oup.com/jeea/article-pdf/9/3/522/10314305/jeea0522.pdf

[42] Finale Doshi-Velez and Been Kim. 2017. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608* (2017).

[43] David Dranove, Daniel Kessler, Mark McClellan, and Mark Satterthwaite. 2003. Is More Information Better? The Effects of "Report Cards" on Health Care Providers. *Journal of Political Economy* 111, 3 (2003), 555–588. doi:10.1086/374180

[44] Carsten Eickhoff. 2018. Cognitive biases in crowdsourcing. In *Proceedings of the eleventh ACM international conference on web search and data mining*. 162–170.

[45] Abdallah El Ali, Karthikeya Puttur Venkatraj, Sophie Morosoli, Laurens Naudts, Natali Helberger, and Pablo Cesar. 2024. Transparent AI disclosure obligations: Who, what, when, where, why, how. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*. 1–11.

[46] Glenn Ellison and Sara Fisher Ellison. 2009. Search, Obfuscation, and Price Elasticities on the Internet. *Econometrica* 77, 2 (2009), 427–452. doi:10.3982/ECTA5708

[47] Alexander Erlei, Abhinav Sharma, and Ujwal Gadiraju. 2024. Understanding choice independence and error types in human-ai collaboration. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–19.

[48] Seyedehdelaram Esfahani, Giovanni De Toni, Bruno Lepri, Andrea Passerini, Katya Tentori, and Massimo Zancanaro. 2024. Preference Elicitation in Interactive and User-centered Algorithmic Recourse: an Initial Exploration. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization* (Cagliari, Italy) *(UMAP '24)*. Association for Computing Machinery, New York, NY, USA, 249–254. doi:10.1145/3627043.3659556

[49] Raymond Fok and Daniel S Weld. 2024. In search of verifiability: Explanations rarely enable complementary performance in AI-advised decision making. *AI Magazine* 45, 3 (2024), 317–332.

[50] Pantea Foroudi, Reza Marvi, and Dongmei Zha. 2025. AI sensation and engagement: Unpacking the sensory experience in human-AI interaction. *International Journal of Information Management* 84 (2025), 102918. doi:10.1016/j.ijinfomgt.2025.102918

[51] Colin Foster and Paul Renie. 2024. Changes in students' confidence calibration across a sequence of low-stakes confidence assessments. *Asian Journal for Mathematics Education* 3, 4 (2024), 406–427. doi:10.1177/27527263241298968 arXiv:https://doi.org/10.1177/27527263241298968

[52] Thomas Franke, Christiane Attig, and Daniel Wessel. 2019. A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *Int. J. Hum. Comput. Interact.* 35, 6 (April 2019), 456–467.

[53] Xavier Gabaix and David Laibson. 2006. Shrouded Attributes, Consumer Myopia, and Information Suppression. *Quarterly Journal of Economics* 121, 2 (2006), 505–540. doi:10.1162/qjec.2006.121.2.505

[54] Ujwal Gadiraju and Agathe Balayn. 2025. The Enterprising and Elusive Prospects of Human-AI Collaboration. In *Enterprise AI*. Springer, 211–243.

[55] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna M. Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for Datasets. *Commun. ACM* 64, 12 (2021), 86–92. doi:10.1145/3458723

[56] Maximilian Germann and Christoph Merkle. 2023. Algorithm aversion in delegated investing. *Journal of Business Economics* 93, 9 (01 Nov 2023), 1691–1727. doi:10.1007/s11573-022-01121-9

[57] Oscar Gomez, Steffen Holter, Jun Yuan, and Enrico Bertini. 2020. ViCE: visual counterfactual explanations for machine learning models. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (Cagliari, Italy) *(IUI '20)*. Association for Computing Machinery, New York, NY, USA, 531–535. doi:10.1145/3377325.3377536

[58] Kimberly Goodyear, Raja Parasuraman, Sergey Chernyak, Ewart de Visser, Poornima Madhavan, Gopikrishna Deshpande, and Frank Krueger. 2017. An fMRI and effective connectivity study investigating miss errors during advice utilization from human and machine agents. *Social neuroscience* 12, 5 (2017), 570–581.

[59] Ben Green and Yiling Chen. 2019. The Principles and Limits of Algorithm-in-the-Loop Decision Making. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 50 (nov 2019), 24 pages. doi:10.1145/3359152

[60] Sanford J Grossman. 1981. The informational role of warranties and private disclosure about product quality. *The Journal of law and Economics* 24, 3 (1981), 461–483.

[61] Gaole He, Nilay Aishwarya, and Ujwal Gadiraju. 2025. Is Conversational XAI All You Need? Human-AI Decision Making With a Conversational XAI Assistant. In *Proceedings of the 30th International Conference on Intelligent User Interfaces (IUI '25)*. Association for Computing Machinery, New York, NY, USA, 907–924. doi:10.1145/3708359.3712133

[62] Gaole He, Abri Bharos, and Ujwal Gadiraju. 2024. To Err Is AI! Debugging as an Intervention to Facilitate Appropriate Reliance on AI Systems. In *Proceedings of the 35th ACM Conference on Hypertext and Social Media* (Poznan, Poland) *(HT '24)*. Association for Computing Machinery, New York, NY, USA, 98–105. doi:10.1145/3648188.3675130

[63] Gaole He, Stefan Buijsman, and Ujwal Gadiraju. 2023. How stated accuracy of an AI system and analogies to explain accuracy affect human reliance on the system. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (2023), 1–29.

[64] Gaole He, Lucie Kuiper, and Ujwal Gadiraju. 2023. Knowing About Knowing: An Illusion of Human Competence Can Hinder Appropriate Reliance on AI Systems. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 113, 18 pages. doi:10.1145/3544548.3581025

[65] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2023. Measures for explainable AI: Explanation goodness, user satisfaction, mental models, curiosity, trust, and human-AI performance. *Frontiers in Computer Science* 5 (2023), 1096257.

[66] Pauline M. Ippolito and Alan D. Mathios. 1990. Information, Advertising, and Health Choices: A Study of the Cereal Market. *RAND Journal of Economics* 21, 3 (1990), 459–480. doi:10.2307/2555457

[67] Alon Jacovi, Ana Marasović, Tim Miller, and Yoav Goldberg. 2021. Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 624–635.

[68] Ginger Zhe Jin. 2018. Artificial intelligence and consumer privacy. In *The economics of artificial intelligence: An agenda*. University of Chicago Press, 439–462.

[69] Ginger Zhe Jin and Andrew Kato. 2006. Price, quality, and reputation: Evidence from an online field experiment. *RAND Journal of Economics* 37, 4 (2006), 983–1005. doi:10.1111/j.1756-2171.2006.tb00067.x

[70] Ginger Zhe Jin and Phillip Leslie. 2003. The Effect of Information on Product Quality: Evidence from Restaurant Hygiene Grade Cards. *Quarterly Journal of Economics* 118, 2 (2003), 409–451. doi:10.1162/003355303321675428

[71] Ginger Zhe Jin and Phillip Leslie. 2009. Reputational Incentives for Restaurant Hygiene. *American Economic Journal: Microeconomics* 1, 1 (2009), 237–267. doi:10.1257/mic.1.1.237

[72] Ginger Zhe Jin, Michael Luca, and Daniel Martin. 2021. Is no news (perceived as) bad news? An experimental investigation of information disclosure. *American Economic Journal: Microeconomics* 13, 2 (2021), 141–173.

[73] Ginger Zhe Jin, Michael Luca, and Daniel Martin. 2022. Complex disclosure. *Management Science* 68, 5 (2022), 3236–3261.

[74] Ginger Zhe Jin and Alan T. Sorensen. 2006. Information and Consumer Choice: The Value of Publicized Health Plan Ratings. *Journal of Health Economics* 25, 2 (2006), 248–275. doi:10.1016/j.jhealeco.2005.06.002

[75] Xiaotong Jin and Jiayang Li. 2025. The Influence of 'Algorithm Aversion' and 'Algorithm Appreciation' Among Consumers in Unstructured Tasks. *International Journal of Consumer Studies* 49, 6 (2025), e70133. doi:10.1111/ijcs.70133 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijcs.70133 e70133 IJC-OA-2025-007.R2.

[76] Patricia K. Kahr, Gerrit Rooks, Martijn C. Willemsen, and Chris C. P. Snijders. 2024. Understanding Trust and Reliance Development in AI Advice: Assessing Model Accuracy, Model Explanations, and Experiences from Previous Interactions. *ACM Trans. Interact. Intell. Syst.* (Aug. 2024). doi:10.1145/3686164 Just

Accepted.

[77] Margot E Kaminski and Gianclaudio Malgieri. 2021. Algorithmic impact assessments under the GDPR: producing multi-layered explanations. *International data privacy law* 11, 2 (2021), 125–144.

[78] Harmanpreet Kaur, Matthew R Conrad, Davis Rule, Cliff Lampe, and Eric Gilbert. 2024. Interpretability gone bad: The role of bounded rationality in how practitioners understand machine learning. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (2024), 1–34.

[79] Simran Kaur, Sara Salimzadeh, and Ujwal Gadiraju. 2026. Incentive-Tuning: Understanding and Designing Incentives for Empirical Human-AI Decision-Making Studies. *arXiv preprint arXiv:2601.15064* (2026).

[80] DeBrae Kennedy-Mayo and Jake Gord. 2025. "Model Cards for Model Reporting" in 2024: Reclassifying Category of Ethical Considerations in Terms of Trustworthiness and Risk Management. In *Future of Information and Communication Conference*. Lecture Notes in Networks and Systems, Vol. 1283. Springer, 179–196. doi:10.1007/978-3-031-84457-7_11

[81] Pranav Khadpe, Ranjay Krishna, Li Fei-Fei, Jeffrey T Hancock, and Michael S Bernstein. 2020. Conceptual metaphors impact perceptions of human-ai collaboration. *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–26.

[82] Rafal Kocielnik, Saleema Amershi, and Paul N Bennett. 2019. Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of ai systems. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–14.

[83] Vivian Lai, Han Liu, and Chenhao Tan. 2020. "Why is 'Chicago' deceptive?" Towards Building Model-Driven Tutorials for Humans. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3313831.3376873

[84] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (Atlanta, GA, USA) *(FAT* '19)*. Association for Computing Machinery, New York, NY, USA, 29–38. doi:10.1145/3287560.3287590

[85] Vivian Lai, Yiming Zhang, Chacha Chen, Q. Vera Liao, and Chenhao Tan. 2023. Selective Explanations: Leveraging Human Input to Align Explainable AI. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 357 (oct 2023), 35 pages. doi:10.1145/3610206

[86] Markus Langer, Tim Hunsicker, Tina Feldkamp, Cornelius J. König, and Nina Grgić-Hlača. 2022. "Look! It's a Computer Program! It's an Algorithm! It's AI!": Does Terminology Affect Human Perceptions and Evaluations of Algorithmic Decision-Making Systems?. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 581, 28 pages. doi:10.1145/3491102.3517527

[87] Johann Laux, Sandra Wachter, and Brent Mittelstadt. 2024. Three pathways for standardisation and ethical disclosure by default under the European Union Artificial Intelligence Act. *Computer Law & Security Review* 53 (2024), 105957.

[88] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[89] Benedikt Leichtmann, Christina Humer, Andreas Hinterreiter, Marc Streit, and Martina Mara. 2023. Effects of Explainable Artificial Intelligence on trust and human behavior in a high-risk decision task. *Computers in Human Behavior* 139 (2023), 107539. doi:10.1016/j.chb.2022.107539

[90] Jingshu Li, Yitian Yang, Q Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025. As Confidence Aligns: Understanding the Effect of AI Confidence on Human Self-confidence in Human-AI Decision Making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. 1–16.

[91] Weixin Liang, Nazneen Rajani, Xinyu Yang, Ezinwanne Ozoani, Eric Wu, Yiqun Chen, Daniel Scott Smith, and James Zou. 2024. Systematic analysis of 32,111 AI model cards characterizes documentation practice in AI. *Nature Machine Intelligence* 6, 7 (2024), 744–753.

[92] Duri Long and Brian Magerko. 2020. What is AI Literacy? Competencies and Design Considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–16. doi:10.1145/3313831.3376727

[93] Luca Longo, Mario Brcic, Federico Cabitza, Jaesik Choi, Roberto Confalonieri, Javier Del Ser, Riccardo Guidotti, Yoichi Hayashi, Francisco Herrera, Andreas Holzinger, Richard Jiang, Hassan Khosravi, Freddy Lecue, Gianclaudio Malgieri, Andrés Páez, Wojciech Samek, Johannes Schneider, Timo Speith, and Simone Stumpf. 2024. Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion* 106 (2024), 102301. doi:10.1016/j.inffus.2024.102301

[94] Shuai Ma, Ying Lei, Xinru Wang, Chengbo Zheng, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2023. Who Should I Trust: AI or Myself? Leveraging Human and AI Correctness Likelihood to Promote Appropriate Trust in AI-Assisted Decision-Making. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany,) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 759, 19 pages. doi:10.1145/3544548.3581058

[95] Shuai Ma, Xinru Wang, Ying Lei, Chuhan Shi, Ming Yin, and Xiaojuan Ma. 2024. "Are You Really Sure?" Understanding the Effects of Human Self-Confidence Calibration in AI-Assisted Decision Making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '24)*. Association for Computing Machinery, New York, NY, USA, Article 840, 20 pages. doi:10.1145/3613904.3642671

[96] Alfred Marshall. 2013. *Principles of economics*. Springer.

[97] Alan D. Mathios. 2000. The Impact of Mandatory Disclosure Laws on Product Choices: An Analysis of the Salad Dressing Market. *Journal of Law and Economics* 43, 2 (2000), 651–678. doi:10.1086/467466

[98] Siddharth Mehrotra, Chadha Degachi, Oleksandra Vereschak, Catholijn M Jonker, and Myrthe L Tielman. 2024. A systematic review on fostering appropriate trust in Human-AI interaction: Trends, opportunities and challenges. *ACM Journal on Responsible Computing* 1, 4 (2024), 1–45.

[99] Daria Mikhaylova, Tommaso Turchi, Gustavo Cevolani, and Alessio Malizia. 2025. Bayesian reasoning for overcoming over-reliance in AI-assisted decision making. (2025).

[100] Paul R. Milgrom. 1981. Good News and Bad News: Representation Theorems and Applications. *The Bell Journal of Economics* 12, 2 (1981), 380–391. doi:10.2307/3003562

[101] Deborah J Miller, Elliot S Spengler, and Paul M Spengler. 2015. A meta-analysis of confidence and judgment accuracy in clinical decision making. *J. Couns. Psychol.* 62, 4 (Oct. 2015), 553–567.

[102] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. doi:10.1016/j.artint.2018.07.007

[103] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*. ACM, 220–229. doi:10.1145/3287560.3287596

[104] Kathleen L Mosier, Linda J Skitka, Susan Heers, and Mark Burdick. 2017. Automation bias: Decision making and performance in high-tech cockpits. In *Decision Making in Aviation*. Routledge, 271–288.

[105] Chris Nosko and Steven Tadelis. 2015. *The limits of reputation in platform markets: An empirical analysis and field experiment*. Technical Report. National Bureau of Economic Research.

[106] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable AI systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*. 340–350.

[107] Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring Bias Affects Mental Model Formation and User Reliance in Explainable AI Systems. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (College Station, TX, USA,) *(IUI '21)*. Association for Computing Machinery, New York, NY, USA, 340–350. doi:10.1145/3397481.3450639

[108] Chioma Ngozi Nwafor, Obumneme Nwafor, and Sanjukta Brahma. 2024. Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. *Scientific Reports* 14, 1 (2024), 25174.

[109] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* 366, 6464 (2019), 447–453.

[110] Myle Ott, Yejin Choi, Claire Cardie, and Jeffrey T. Hancock. 2011. Finding deceptive opinion spam by any stretch of the imagination. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1* (Portland, Oregon) *(HLT '11)*. Association for Computational Linguistics, USA, 309–319.

[111] Niccolò Pescetelli and Nicholas Yeung. 2021. The role of decision confidence in advice-taking and trust formation. *J. Exp. Psychol. Gen.* 150, 3 (March 2021), 507–526.

[112] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI conference on human factors in computing systems*. 1–52.

[113] Jiaming Qu, Jaime Arguello, and Yue Wang. 2025. Understanding the Effects of Explaining Predictive but Unintuitive Features in Human-XAI Interaction. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency (FAccT '25)*. Association for Computing Machinery, New York, NY, USA, 296–311. doi:10.1145/3715275.3732021

[114] Inioluwa Deborah Raji, Andrew Smart, Rebecca N White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 33–44.

[115] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2022. Deciding fast and slow: The role of cognitive biases in ai-assisted decision-making. *Proceedings of the ACM on Human-computer Interaction* 6, CSCW1 (2022), 1–22.

[116] Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. *arXiv preprint arXiv:2507.06306* (2025).

[117] Luis Rayo and Ilya Segal. 2010. Optimal Information Disclosure. *Journal of Political Economy* 118, 5 (2010), 949–987.

[118] Amy Rechkemmer and Ming Yin. 2022. When Confidence Meets Accuracy: Exploring the Effects of Multiple Performance Indicators on Trust in Machine Learning Models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans,LA, USA) *(CHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 535, 14 pages. doi:10.1145/3491102.3501967

[119] Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. 2022. Experimental evidence of effective human–AI collaboration in medical decision-making. *Scientific reports* 12, 1 (2022), 14952.

[120] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. *arXiv preprint arXiv:2005.04118* (2020).

[121] Cynthia Rudin. 2019. Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. *Nature Machine Intelligence* 1 (2019), 206–215. doi:10.1038/s42256-019-0048-x

[122] Sara Salimzadeh and Ujwal Gadiraju. 2024. When in Doubt! Understanding the Role of Task Characteristics on Peer Decision-Making with AI Assistance. In *Proceedings of the 32nd ACM Conference on User Modeling, Adaptation and Personalization*. 89–101.

[123] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2023. A missing piece in the puzzle: Considering the role of task complexity in human-ai decision making. In *Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*. 215–227.

[124] Sara Salimzadeh, Gaole He, and Ujwal Gadiraju. 2024. Dealing with uncertainty: Understanding the impact of prognostic versus diagnostic tasks on trust and reliance in human-AI decision making. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. 1–17.

[125] Jakob Schoeffer, Niklas Kuehl, and Yvette Machowski. 2022. "There Is Not Enough Information": On the Effects of Explanations on Perceptions of Informational Fairness and Trustworthiness in Automated Decision-Making. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1616–1628. doi:10.1145/3531146.3533218

[126] Gian Luca Scoccia, Marco Autili, Giovanni Stilo, and Paola Inverardi. 2022. An empirical study of privacy labels on the Apple iOS mobile app store. In *Proceedings of the 9th IEEE/ACM International Conference on Mobile Software Engineering and Systems*. 114–124.

[127] Carl Shapiro. 1983. Premiums for high quality products as returns to reputations. *The quarterly journal of economics* 98, 4 (1983), 659–679.

[128] Hong Shen, Leijie Wang, Wesley H. Deng, Ciell Brusse, Ronald Velgersdijk, and Haiyi Zhu. 2022. The Model Card Authoring Toolkit: Toward Community-centered, Deliberation-driven AI Design. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. ACM. doi:10.1145/3531146.3533110

[129] Ben Shneiderman. 2022. *Human-centered AI*. Oxford University Press.

[130] Telmo Silva Filho, Hao Song, Miquel Perello-Nieto, Raul Santos-Rodriguez, Meelis Kull, and Peter Flach. 2023. Classifier calibration: a survey on how to assess and improve predicted class probabilities. *Machine Learning* 112, 9 (01 Sep 2023), 3211–3260. doi:10.1007/s10994-023-06336-7

[131] Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. 2025. Trust Me, I'm Wrong: High-Certainty Hallucinations in LLMs. *arXiv preprint arXiv:2502.12964* (2025).

[132] Michael Spence. 1977. Consumer misperceptions, product failure and producer liability. *The Review of Economic Studies* 44, 3 (1977), 561–572.

[133] Aaron Springer and Steve Whittaker. 2019. Progressive disclosure: empirically motivated approaches to designing effective transparency. In *Proceedings of the 24th international conference on intelligent user interfaces*. 107–120.

[134] Constantine Stephanidis, Gavriel Salvendy, Margherita Antona, Vincent G Duffy, Qin Gao, Waldemar Karwowski, Shin'ichi Konomi, Fiona Nah, Stavroula Ntoa, Pei-Luen Patrick Rau, et al. 2025. Seven HCI grand challenges revisited: Five-year progress. *International Journal of Human–Computer Interaction* (2025), 1–49.

[135] Mark Steyvers, Heliodoro Tejeda, Aakriti Kumar, Catarina Belem, Sheer Karny, Xinyue Hu, Lukas W Mayer, and Padhraic Smyth. 2025. What large language models know and what people think they know. *Nature Machine Intelligence* 7, 2 (2025), 221–231.

[136] Julia Stoyanovich and Bill Howe. 2019. Nutritional labels for data and models. *A Quarterly bulletin of the Computer Society of the IEEE Technical Committee on Data Engineering* 42, 3 (2019).

[137] Siddharth Swaroop, Zana Buçinca, Krzysztof Z Gajos, and Finale Doshi-Velez. 2024. Accuracy-time tradeoffs in AI-assisted decision making under time pressure. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*. 138–154.

[138] Maxwell Szymanski, Vero Vanden Abeele, and Katrien Verbert. 2024. Designing and Evaluating Explanations for a Predictive Health Dashboard: A User-Centred Case Study. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 514, 8 pages. doi:10.1145/3613905.3637140

[139] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, Aimilios Lallas, Jan Lapins, Caterina Longo, Josep Malvehy, Michael A Marchetti, Ashfaq Marghoob, Scott Menzies, Amanda Oakley, John Paoli, Susana Puig, Christoph Rinner, Cliff Rosendahl, Alon Scope, Christoph Sinz, H. Peter Soyer, Luc Thomas, Iris Zalaudek, and Harald Kittler. 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *Lancet Oncology* 20, 7 (2019), 938–947. doi:10.1016/S1470-2045(19)30333-X

[140] Philipp Tschandl, Christoph Rinner, Zoi Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli, Susana Puig, Cliff Rosendahl, H. Peter Soyer, Iris Zalaudek, and Harald Kittler. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26 (2020), 1229–1234. doi:10.1038/s41591-020-0942-0

[141] Niels van Berkel, Jorge Goncalves, Daniel Russo, Simo Hosio, and Mikael B. Skov. 2021. Effect of Information Presentation on Fairness Perceptions of Machine Learning Predictors. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 245, 13 pages. doi:10.1145/3411764.3445365

[142] Colin Van Noordt and Gianluca Misuraca. 2022. Artificial intelligence for the public sector: results of landscaping the use of AI in government across the European Union. *Government information quarterly* 39, 3 (2022), 101714.

[143] Helena Vasconcelos, Matthew Jörke, Madeleine Grunde-McLaughlin, Tobias Gerstenberg, Michael S Bernstein, and Ranjay Krishna. 2023. Explanations can reduce overreliance on ai systems during decision-making. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW1 (2023), 1–38.

[144] Xinru Wang, Chen Liang, and Ming Yin. 2023. The effects of AI biases and explanations on human decision fairness: a case study of bidding in rental housing markets. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence* (Macao, P.R.China) *(IJCAI '23)*. Article 343, 9 pages. doi:10.24963/ijcai.2023/343

[145] Siyuan Yan, Zhen Yu, Clare Primiero, Cristina Vico-Alonso, Zhonghua Wang, Litao Yang, Philipp Tschandl, Ming Hu, Lie Ju, Gin Tan, Vincent Tang, Aik Beng Ng, David Powell, Paul Bonnington, Simon See, et al. 2025. A multimodal vision foundation model for clinical dermatology. *Nature Medicine* (2025). doi:10.1038/s41591-025-03747-y

[146] Wenli Yang, Yuchen Wei, Hanyu Wei, Yanyu Chen, and Guan Huang. 2023. Survey on Explainable AI: From Approaches, Limitations and Applications Aspects. *Human-Centric Intelligent Systems* (2023).

[147] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the Effect of Accuracy on Trust in Machine Learning Models. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) *(CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. doi:10.1145/3290605.3300509

[148] Mireia Yurrita, Tim Draws, Agathe Balayn, Dave Murray-Rust, Nava Tintarev, and Alessandro Bozzon. 2023. Disentangling Fairness Perceptions in Algorithmic Decision-Making: the Effects of Explanations, Human Oversight, and Contestability. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 134, 21 pages. doi:10.1145/3544548.3581161

[149] Mireia Yurrita, Himanshu Verma, Agathe Balayn, Ujwal Gadiraju, Sylvia C. Pont, and Alessandro Bozzon. 2025. Towards Effective Human Intervention in Algorithmic Decision-Making: Understanding the Effect of Decision-Makers' Configuration on Decision-Subjects' Fairness Perceptions. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 1028, 21 pages. doi:10.1145/3706598.3713145

[150] Daochen Zha, Zaid Pervaiz Bhat, Kwei-Herng Lai, Fan Yang, Zhimeng Jiang, Shaochen Zhong, and Xia Hu. 2025. Data-centric Artificial Intelligence: A Survey. *ACM Comput. Surv.* 57, 5, Article 129 (Jan. 2025), 42 pages. doi:10.1145/3711118

[151] Yunfeng Zhang, Q. Vera Liao, and Rachel K. E. Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency* (Barcelona, Spain) *(FAT* '20)*. Association for Computing Machinery, New York, NY, USA, 295–305. doi:10.1145/3351095.3372852

Appendix

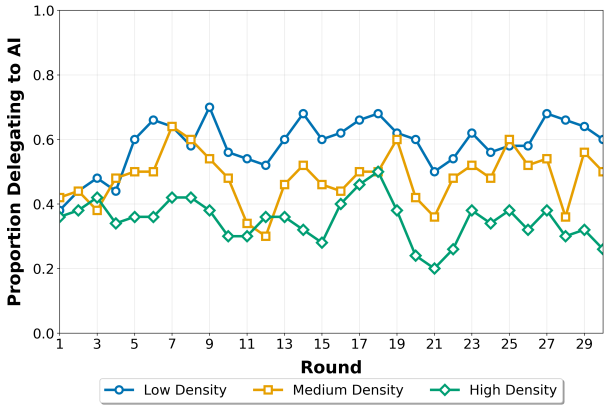

Figure 5: Average subject delegation in `No Disclosure` between density conditions.

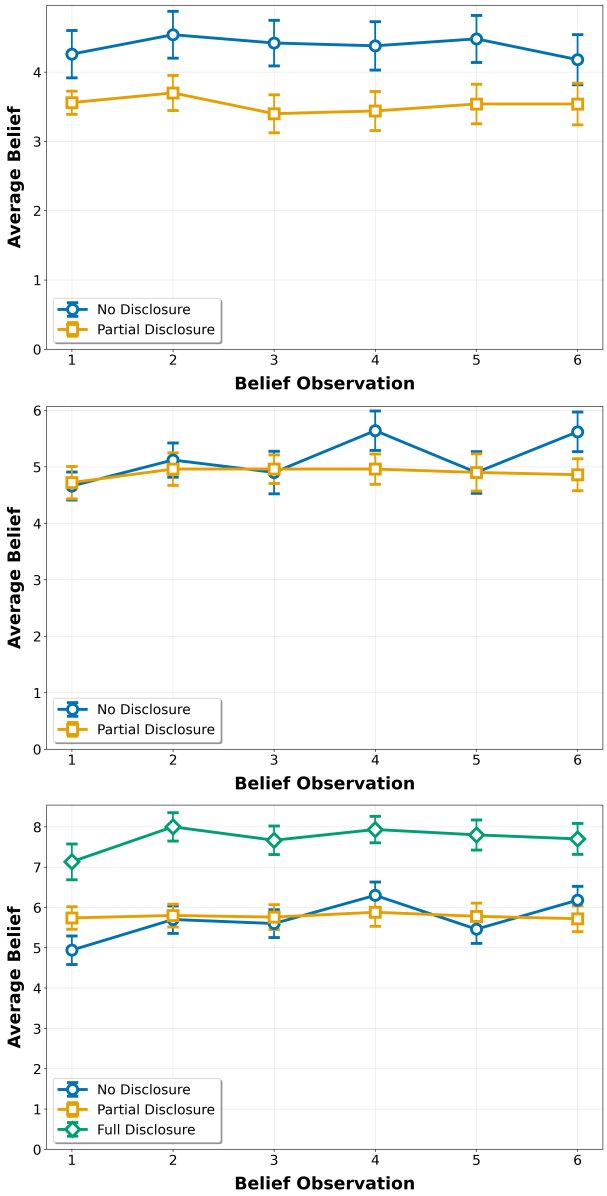

Figure 6: Belief evolution over time in order `Low Density`, `Medium Density`, `High Density`.

Table 5: Summary of participants' task familiarity across conditions. Mean (M) and standard deviation (SD) are reported for each of the three tasks (loan approval, deceptive review detection, and skin cancer prediction). *Sample* **indicates the number of observations per condition.**

Condition	Sample	Loan_{M}	Loan_{SD}	$\text{Reviews}_{\text{M}}$	$\text{Reviews}_{\text{SD}}$	Cancer_{M}	$\text{Cancer}_{\text{SD}}$
Full disclosure, high density	30	2.43	1.36	2.03	1.00	1.60	0.89
No disclosure, low density	50	2.40	1.41	2.10	1.20	1.78	1.13
No disclosure, medium density	50	2.32	1.19	1.90	1.05	1.50	0.74
No disclosure, high density	50	2.42	1.18	2.00	1.21	1.60	0.97
Partial disclosure, low density	50	2.46	1.18	2.12	1.21	1.76	0.87
Partial disclosure, medium density	50	2.36	1.17	2.14	1.16	1.56	0.93
Partial disclosure, high density	50	2.06	1.11	1.86	0.97	1.58	0.99

Notes. We assessed the distribution of task familiarity scores for loan approval, deceptive reviews, and skin cancer tasks using the Shapiro–Wilk normality test. All three tasks significantly deviated from normality (loan approval: $W = 0.871$, $p < .001$; deceptive reviews: $W = 0.819$, $p < .001$; skin cancer: $W = 0.699$, $p < .001$). Consequently, we applied non-parametric Kruskal–Wallis tests to examine potential differences across conditions. Results indicated no significant differences in task familiarity across conditions for any of the tasks (loan approval: $\chi^2 = 3.64$, $df = 6$, $p = 0.72$; deceptive reviews: $\chi^2 = 2.30$, $df = 6$, $p = 0.89$; skin cancer: $\chi^2 = 3.89$, $df = 6$, $p = 0.69$).

Table 6: Time trend in delegation under `No Disclosure`: mixed-effects logit with subject RE, task FE, and density×round interaction.

	Delegation Share (Log Odds)
Density (vs Low)	
Medium	−0.277 (0.359)
High	-0.743^{*} (0.359)
Round (1–30)	
Round (centered at 1)	0.025^{***} (0.007)
Medium × Round	−0.020 (0.010)
High × Round	-0.040^{***} (0.010)
Task FE (domain)	Yes
Random intercept SD (user)	1.563
Observations	4,500
Users (groups)	150

Notes. Coefficients are log-odds; standard errors in parentheses. Model: GLMM (logit) with random intercept for subject (1 | user_id) and task fixed effects (domain). Baselines: density=*Low*; task=*cancer_prediction*. Time is a continuous round index from 1–30, centered at 1. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 7: Delegation (logit, subject RE) with disclosure × performance: *Cancer Prediction*. Baselines: No disclosure and Underperform.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	−0.114 (0.513)	−0.317 (0.496)	−0.466 (0.881)	−0.470 (0.938)
Full	—	—	—	−1.211 (0.893)
Performance (vs Underperform)				
Outperform	0.853 (0.677)	−0.551 (0.499)	−1.852* (0.795)	−1.898* (0.845)
Interactions				
Partial × Outperform	−0.538 (2.034)	0.634 (0.712)	0.512 (0.934)	0.522 (0.994)
Full × Outperform	—	—	—	4.373*** (1.263)
Time FE (round)	Yes	Yes	Yes	Yes
Random intercept SD (user)	1.759	1.307	1.126	1.221
Observations	720	830	960	1,150
Users (groups)	72	83	96	115
AIC	832.7	1,012.7	1,034.9	1,241.9

Notes. Mixed-effects logit with random intercept for subject (1|`user_id`) and time fixed effects (round). Coefficients are log-odds; standard errors in parentheses. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 8: Delegation (logit, subject RE) with disclosure × performance: *Deceptive Hotel Reviews*. Baselines: No disclosure and Underperform.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	−0.630 (0.368)	−0.257 (0.409)	−1.334* (0.614)	−1.374* (0.684)
Full	—	—	—	−2.114** (0.658)
Performance (vs Underperform)				
Outperform	−0.996* (0.505)	−0.226 (0.444)	−2.471*** (0.564)	−2.560*** (0.626)
Interactions				
Partial × Outperform	1.161 (1.018)	0.379 (0.713)	1.115 (0.657)	1.161 (0.733)
Full × Outperform	—	—	—	4.802*** (0.916)
Time FE (round)	Yes	Yes	Yes	Yes
Random intercept SD (user)	1.231	1.269	0.719	0.874
Observations	760	870	960	1,180
Users (groups)	76	87	96	118
AIC	952.4	1,084.2	1,113.8	1,363.6

Notes. Mixed-effects logit with random intercept for subject (1|`user_id`) and time fixed effects (round). Coefficients are log-odds; standard errors in parentheses. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 9: Delegation (logit, subject RE) with disclosure × performance: *Loan Prediction*. Baselines: No disclosure and Underperform.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Disclosure (vs No)				
Partial	−0.723 (0.430)	−0.688 (0.484)	−0.610 (0.470)	−0.624 (0.522)
Full	—	—	—	−1.225** (0.447)
Performance (vs Underperform)				
Outperform	−0.251 (0.600)	−1.012* (0.496)	−1.339*** (0.381)	−1.372** (0.422)
Interactions				
Partial × Outperform	1.306 (1.346)	0.665 (0.710)	0.325 (0.522)	0.345 (0.579)
Full × Outperform	—	—	—	—
Time FE (round)	Yes	Yes	Yes	Yes
Random intercept SD (user)	1.432	1.312	0.683	0.820
Observations	700	870	960	1,190
Users (groups)	70	87	96	119
AIC	854.3	1,064.8	1,164.0	1,440.5

Notes. Mixed-effects logit with random intercept for subject (1|`user_id`) and time fixed effects (round). Coefficients are log-odds; standard errors in parentheses. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 10: Always delegate (OLS with subject-clustered SEs). Dependent variable: indicator equals 1 if a participant always delegated within a task. Baseline disclosure is *No*.

	Low density (No vs Partial)	**Medium density** (No vs Partial)	**High density** (No vs Partial)	**High density** (No vs Partial vs Full)
Partial disclosure	0.160* (0.075)	−0.007 (0.057)	−0.027 (0.029)	−0.027 (0.029)
Full disclosure	—	—	—	0.236** (0.077)
Task indicators (baseline: Cancer Prediction)				
Hotel Reviews	−0.040 (0.043)	−0.040 (0.040)	0.000 (0.020)	−0.023 (0.020)
Loan Prediction	0.020 (0.035)	−0.040 (0.040)	0.000 (0.025)	−0.031 (0.027)
Task FE (domain)	Yes	Yes	Yes	Yes
Clustered SE (user)	Yes	Yes	Yes	Yes
Observations (N)	300	300	300	390

Notes. OLS with standard errors clustered at the subject level. Coefficients with standard errors in parentheses. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 11: Mixed-effects logistic regression of delegation (log-odds). Random intercept by subject; task and time fixed effects included. Baseline treatment is *No* × *Low*.

	No/Partial × Low/Med/High	No/Partial × Low/Med/High + Full × High
Treatment (vs No×Low)		
No × Medium	−0.703* (0.329)	−0.712* (0.346)
No × High	−1.371*** (0.329)	−1.395*** (0.346)
Partial × Low	0.438 (0.334)	0.454 (0.351)
Partial × Medium	−0.784* (0.328)	−0.798* (0.345)
Partial × High	−1.447*** (0.326)	−1.462*** (0.343)
Full × High	—	0.083 (0.405)
Controls (standardized)		
Affinity for technology	−0.282** (0.103)	−0.270** (0.104)
AI literacy	−0.158 (0.111)	−0.174 (0.113)
Risk	−0.189 (0.113)	−0.216 (0.115)
Task familiarity	−0.092* (0.046)	−0.113* (0.044)
Task fixed effects (domain)	Yes	Yes
Time fixed effects (round)	Yes	Yes
Random intercept SD (user)	1.557	1.644
Observations	9,000	9,900
Users (groups)	300	330
AIC	9,763.9	10,597.4

Notes. Coefficients are log-odds; robust standard errors in parentheses. Models: GLMM (logit) with random intercept for subject (1|user_id), task fixed effects (domain), and time fixed effects (round). Controls are z-scored user means for affinity and AI literacy, z-scored risk, and a z-scored task-specific familiarity mapped to the row's task. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 12: OLS regressions of total task coins. Cluster-robust SEs (subjects) in parentheses. Baseline treatment is *No* × *Low*.

	(1) FE + controls	(2) + performance & *AI*-share interactions
Treatments (vs no_low)		
no_medium	−39.129*** (5.819)	4.034 (7.947)
no_high	−57.506*** (6.985)	12.860 (8.074)
partial_low	35.792*** (6.729)	9.033 (7.845)
partial_medium	−14.210** (5.424)	2.620 (8.193)
partial_high	−50.719*** (5.601)	13.770 (7.578)
full_high	35.014*** (9.441)	—
Performance / Delegation		
AI share	—	72.308*** (14.469)
Own performance	—	123.545*** (7.434)
no_medium × AI share	—	−74.913*** (20.191)
no_high × AI share	—	−165.658*** (20.865)
partial_low × AI share	—	43.889* (17.896)
partial_medium × AI share	—	−21.433 (20.685)
partial_high × AI share	—	−150.888*** (19.225)
Controls (z-scored)		
Affinity for technology	−4.482* (2.130)	−2.424 (1.332)
AI literacy	4.441* (2.184)	1.761 (1.572)
Risk	−6.738** (2.249)	−0.385 (1.400)
Task familiarity	−3.975 (2.237)	0.351 (1.412)
Task FE (domain)	Yes	Yes
SEs	Clustered by user	Clustered by user

Notes. Column (1): OLS with treatment fixed effects and controls. Column (2): adds own performance and *AI*-delegation share with treatment × AI-share interactions; the full_high cell is excluded in this six-cell specification. Significance: $^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$.

Table 13: Post-hoc pairwise comparisons using Dunn test with Bonferroni correction for the number of AI hovered per condition ($^{*}p<.05$, $^{**}p<.01$, $^{***}p<.001$).

Comparison	Z	p_raw	p_adj
full_high - partial_high	3.0829	.0020	.0123*
full_high - partial_low	-0.7411	.4586	1.0000
partial_high - partial_low	-4.3568	.0000	.0001***
full_high - partial_medium	0.3352	.7375	1.0000
partial_high - partial_medium	-3.1553	.0016	.0096**
partial_low - partial_medium	1.2288	.2192	1.0000

To obtain further insights about participants' behaviors, we explored the average numbers of AIs they hovered during the study, considering all the **`Partial Disclosure`** conditions and the **`Full Disclosure`** benchmark (see Table 13 in the Appendix). Preliminary analysis indicated that the data were not normally distributed, as demonstrated by the Shapiro-Wilk normality test ($W = 0.916$, $p < .001$). Therefore, we employed the non-parametric Kruskal-Wallis test, which revealed significant differences between conditions ($\chi^2 = 21.5$, $df = 3$, $p < .001$). Based on this result, post-hoc pairwise comparisons were conducted using the Dunn test with Bonferroni correction for multiple comparisons. The comparisons showed that *partial_high* had significantly higher AIs hovering from *partial_low* ($Z = 4.36$, $p_{adj} < .001$) and from *partial_medium* ($Z = 3.16$, $p_{adj} = .0096$). Additionally, *full_high* conditions had significantly less AIs hovered from *partial_high* ($Z = -3.08$, $p_{adj} = .0123$).