

Designing Human-AI Collaboration: A Sufficient-Statistic Approach

Nikhil Agarwal, Alex Moehring, Alexander Wolitzky *

April 18, 2025

Abstract

We develop a sufficient-statistic approach to designing collaborative human-AI decision-making policies in classification problems, where AI predictions can be used to either automate decisions or selectively assist humans. The approach allows for endogenous and biased beliefs, and effort crowd-out, without imposing a structural model of human decision-making. We deploy and validate our approach in an online fact-checking experiment. We find that humans under-respond to AI predictions and reduce effort when presented with confident AI predictions. AI under-response stems more from human overconfidence in own-signal precision than from under-confidence in AI. The optimal policy automates cases where the AI is confident and delegates uncertain cases to humans while fully disclosing the AI prediction. Although automation is valuable, the additional benefit from assisting humans with AI predictions is negligible.

JEL: C91, D83, D89, D47.

Keywords: Artificial Intelligence, Human-AI Interaction, Belief Updating, Information Design, Fact-Checking.

*Agarwal: Department of Economics, MIT and NBER, email: agarwaln@mit.edu. Moehring: Daniels School of Business, Purdue University, email: moehring@purdue.edu. Wolitzky: Department of Economics, MIT, email: wolitzky@mit.edu. Ray Huang, Bobby Upton, and Crystal Qian provided invaluable research assistance. We are grateful to David Atkin, Glenn Ellison, Drew Fudenberg, Parag Pathak, Frank Schilbach, and Tomasz Strzalecki for valuable comments. We are particularly grateful to Tobias Salz for initial discussions on the project. The authors acknowledge support from the Alfred P. Sloan Foundation (2022-17182). The experiment was pre-registered on the AEA registry, number AEARCTR-0013990. The preanalysis plans are available at www.socialscienceregistry.org/trials/13990.

1 Introduction

The performance of Artificial Intelligence tools has improved rapidly in recent years (Maslej et al., 2024), with many predictive tools matching or surpassing humans (Kleinberg et al., 2017; Agrawal et al., 2018; Lai et al., 2021). Correspondingly, there has been great interest in how AI assistance affects human performance (Noy and Zhang, 2023; Brynjolfsson et al., 2025) and in the design of human-AI collaborative systems that consider which specific cases or tasks to automate or to assign to humans, either with or without AI assistance (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023).

A challenge in designing human-AI collaboration is that the space of possible collaborative designs is large, and it can be difficult to predict how humans will respond to different designs. In particular, humans can exhibit biases in belief updating in response to AI information (Agarwal et al., 2023), and AI information can crowd out human effort in acquiring or processing information—phenomena known as algorithmic aversion (Dietvorst et al., 2015), automation bias (Skitka et al., 1999), or “falling asleep at the wheel” (Dell’Acqua, 2022). The complexity of these possible responses, together with the high dimensionality of the space of possible collaborative designs, frustrates the search for an optimal design via experimentation or structural modeling.

This paper develops a sufficient-statistic approach for designing human-AI collaboration for binary classification tasks, where each of several cases must receive a classification $a \in \{0, 1\}$.¹ The sufficient statistic, $V(x)$, is the probability that a human decision-maker correctly classifies a case when she observes a calibrated AI assessment that the probability that the correct classification is 1 is $x \in [0, 1]$.² We allow any AI system that selectively automates classification tasks based on its assessment and/or delegates tasks to a human decision-maker while disclosing a (potentially imperfect) signal of its assessment. Under the maintained assumption that the function V does not depend on the information disclosure policy, results from the literature on information design (Dworczak and Martini, 2019) imply that V can be used to find the optimal design in this space. That is, conditional on V , the design problem does not depend on any other aspects of the human-AI interaction, such as humans’ information, behavioral biases, or effort responses.

The sufficient-statistic approach has important advantages over two natural alternatives. One alternative estimates a fully-specified structural model of human behavior and belief

¹We focus on binary classification problems where the objective is maximizing accuracy. However, our approach extends to multi-class classification problems and alternative objectives.

²In our setting, the sufficient statistic is a function rather than a number as is usually the case in public finance (Chetty, 2009).

updating and solves for the optimal design. This approach requires stronger behavioral assumptions, and estimating such a model is likely to require similar data. A second alternative experimentally tests a range of designs to find the optimal one. However, this approach is impractical because the space of potential designs is large. Moreover, there is no guarantee that the highest-performing design tested is globally optimal.

We implement and validate our approach in an incentivized online experiment on fact-checking, where participants are tasked with checking the veracity of statements. Fact-checking is an important setting for studying human-AI collaboration because the veracity of public statements is of great concern (Lazer et al., 2018), and both human and AI fact-checkers are widely employed. While media outlets, independent fact-checking organizations, and digital platforms have long relied on professional human fact-checkers (International Fact-Checking Network, 2023; YouTube, Accessed February 7, 2025; Facebook, Accessed August 12, 2024), the growth in the number of statements to be checked has led to interest in using laypeople for fact-checking (Allen et al., 2021; X Community Notes, 2025; Kaplan, 2025), as well as in fully automated fact-checking (Guo et al., 2022; International Fact-Checking Network, 2023). Understanding how to design AI tools and human-AI collaborative systems to improve fact-checking is thus an ongoing practical challenge.

Besides being an important setting for human-AI collaboration, fact-checking is also convenient for experimental purposes. Fact-checking is easy to explain and can be conducted by untrained experimental participants. Measuring accuracy in fact-checking is straightforward, as there are established databases of true and false statements with curated ground-truth labels, such as the FEVEROUS database (Aly et al., 2021), which we use in our experiment. Finally, fact-checking is representative of other binary classification tasks, such as medical diagnosis (Chan et al., 2022; Agarwal et al., 2023), judicial bail decisions (Kleinberg et al., 2017), and resume screening (Li et al., 2020).

Our experiment proceeds in two stages. The first stage estimates the sufficient statistic V by measuring classification accuracy under varying AI predictions and solves for optimal and approximately optimal designs. We consider designs where automation is allowed as well as designs where humans must make the classification decision, as in many settings—potentially including fact-checking—there may be a societal preference for humans to make final decisions.³ In the second stage, we implement five designs derived from the first-stage estimates in a within-participant experimental design, and test the sufficient-statistic approach by comparing the predicted classification accuracy from our first-stage estimates against the

³For example, the algorithmic aversion literature finds that humans often prefer human decisions over more accurate algorithmic decisions (Dietvorst et al., 2015; Longoni et al., 2019)

second-stage experimental results.

The first-stage experimental results yield several insights and predictions. First, the estimated function V is convex. This property implies that fully disclosing the AI prediction is optimal for all cases that are delegated to human decision-makers. This finding contrasts with prior theoretical and empirical results (Athey et al., 2020; Dell’Acqua, 2022) that find that partial disclosure of AI information can be optimal because disclosing more precise information crowds out human effort in information acquisition. While we also find effort crowding-out, this effect is too weak to overturn the direct benefit of providing more precise AI information.

Second, when the disclosed AI assessment is confident (x is close to 0 or 1), humans’ classification accuracy $V(x)$ is significantly lower than the accuracy under automation, which equals $\max\{x, 1 - x\}$. This implies that humans under-respond to the AI assessment when updating their beliefs, because simply following the AI prediction would increase accuracy whenever $V(x) < \max\{x, 1 - x\}$.⁴ It also implies that automating these cases is optimal.

Third, because uncertain AI predictions add little value to humans’ own predictions, we predict that a policy that automates cases where the AI is confident and delegates cases where the AI is uncertain to humans without AI assistance is approximately optimal. Thus, while both humans and AI add value, the value of human-AI *collaboration*—rather than selective automation and delegation—is negligible.⁵

To summarize, the first-stage results predict that the optimal design automates cases where the AI is confident and delegates the remaining cases to humans while providing them with the AI assessment. We call this policy **Full Disclosure + Automation (FDA)**. Accuracy under FDA is predicted to be similar to that under **No Disclosure + Automation (NDA)**, where cases where the AI is confident are automated, and the rest are delegated to humans without AI assistance. In addition, we predict that the optimal design when automation is infeasible is **Full Disclosure + No Automation (FDNA)**, where humans are provided with the AI assessment. This design is predicted to significantly outperform **No Disclosure + No Automation (NDNA)**, where humans do not receive AI assistance. Finally, we also predict that accuracy under FDNA is very similar to that under a simpler **Stoplight (SL)** policy, where the AI communicates only one of three possible signals (e.g., “Likely False,” “Uncertain,” “Likely True,” or “Red,” “Yellow,” “Green”).

The second stage experiment tests whether the sufficient-statistic approach accurately predicts the performance of these five policies. All predictions are within 1.6 percentage

⁴Under-response to information is a common finding in behavioral economics (Benjamin, 2019).

⁵However, our result differs from the finding in Agarwal et al. (2023) that humans assisted with uncertain AI predictions perform worse than unassisted humans.

points of experimental estimates, and the differences are not significant at the 1% level. In addition, the qualitative predictions are all borne out: FDA is the best policy when automation is feasible but is statistically indistinguishable from NDA; and FDNA is the best policy when automation is infeasible but is indistinguishable from SL, while NDNA is significantly worse.⁶ These results suggest that the sufficient statistic assumption is a good guide for designing human-AI collaboration in our context.

In addition to designing human-AI collaboration using the sufficient statistic V , we also analyze the mechanisms that determine the shape of this function (and hence the optimal designs and their accuracy). In particular, we decompose the impact of behavioral biases and effort response to AI information.

We first estimate a sharp lower bound for the impact of human under-response to AI on accuracy. The bound is obtained by estimating the accuracy of an optimal classifier based on both AI predictions and humans' reported probability assessments. We find that at least 7.7% of incorrect classifications humans make with AI assistance are attributable to errors in belief updating. We also find that the optimal FDA policy approximately achieves the optimal classifier benchmark. This implies that there is little benefit to considering richer collaborative designs where humans' probability assessments can be communicated to the AI.

We next examine whether humans under-respond to AI information because they are overconfident in the accuracy of their own information or under-confident in the AI. To do so, we estimate the distribution of our participants' private information and the update rule participants use to combine their signals with AI predictions. Our method first identifies the distribution of participants' signals s conditional on observed effort (as well as the state and the AI assessment) using their reports in the NDNA treatment. We then use the observed effort distribution in the FDNA treatment to calculate the implied signal distribution in this treatment. Finally, we estimate the update rule $p(s, x)$ to fit the observed reports in FDNA.⁷ We find that AI under-response is almost entirely due to overconfidence in own-signal precision: humans' beliefs are too sensitive to their own signals relative to a Bayesian benchmark but are appropriately sensitive to AI predictions. This result contrasts starkly with prior work that attributes AI under-response to under-confidence in AI signal precision (Agarwal et al., 2023).

Finally, we find that providing accurate AI information crowds out human effort, but the

⁶More precisely, estimated accuracy under FDNA is 0.2 percentage points below that under SL.

⁷The estimated model assumes that our participants use a common update rule and that their signal distribution depends only on effort, the underlying state, and the AI assessment, and not directly on the disclosed AI assessment conditional on these variables.

impact of this effect on the precision of humans’ signals is small.

Related Literature

Comparing predictive AI tools and human decisions is an active area of research (Kleinberg et al., 2017; Mullainathan and Obermeyer, 2022). Several papers compare the accuracy of humans with AI assistance to either humans or AI alone (Angelova et al., 2023; Agarwal et al., 2023; Vaccaro et al., 2024).⁸ Rather than comparing humans and AI, we develop an approach to optimally designing human-AI collaborative systems.

Our focus on data-driven design of human-AI collaboration is shared with the “algorithmic triage” problem in computer science (e.g. Mozannar and Sontag (2020)) and with Raghu et al. (2019) and Agarwal et al. (2023) in economics. We highlight two key differences. First, these papers abstract away from endogenous changes in human beliefs or effort in response to the set of cases that are delegated or automated. However, effort crowding-out has been found to be an important consideration when humans use AI tools (Athey et al., 2020; Dell’Acqua, 2022), and we argue in Appendix D.1 that endogenous belief responses are similarly important in our setting. Second, optimal collaboration design using these earlier approaches requires direct experimentation, because these approaches lack a model for predicting accuracy under counterfactual AI assessments. In addition, none of these papers tests the performance of the optimal policy in a second-stage experiment.

Our sufficient-statistic approach for predicting accuracy in counterfactual policies builds on insights from information design (Kamenica and Gentzkow, 2011). Our sufficient statistic $V(x)$ is the designer’s indirect utility from inducing a posterior mean assessment x , as in Dworzak and Martini (2019). Arieli et al. (2023) notes that this “mean-measurable” design problem arises when the designer discloses information about a signal of an underlying binary state; we extend this observation to the case where the decision-maker additionally observes a private signal that is measurable with respect to x and the underlying state. Like us, De Clippel and Zhang (2022) studies information design with a non-Bayesian receiver. Rather than estimating the designer’s biased belief-updating function (which is a model primitive in De Clippel and Zhang (2022)), we estimate the designer’s indirect utility $V(x)$ and apply standard information design arguments. Finally, a growing experimental literature tests the assumptions and predictions of information design (e.g., Fréchette et al. (2022)), rather than using it for optimal design. In addition to these differences, to our knowledge, this paper is the first to apply information design techniques to human-AI collaboration.

⁸A strand of this literature considers heterogeneous effects by baseline characteristics (Yu et al., 2024).

Some of our empirical results parallel findings in prior experiments on biased belief updating. For instance, under-response to new information is a common finding in behavioral economics (Benjamin, 2019).⁹ We replicate this finding but also go beyond it by decomposing under-response to new information into overconfidence in own-signal precision and under-confidence in the precision of the new information, finding that in our setting under-response is driven almost entirely by the first effect. To do so, we offer a novel definition of over- or under-response to an information source (related to Augenblick et al. (2025)).

A unique feature of our study is the use of a two-stage experiment, where the first stage estimates a sufficient statistic that is used to design an optimal policy, and the second stage validates the design. We are aware of only a handful of papers in economics that design an optimal policy and test it in a second-stage experiment, including Misra and Nair (2011), Dubé and Misra (2023), and Ostrovsky and Schwarz (2023). Our approach is closest to Ostrovsky and Schwarz (2023), who use insights from auction theory to derive a sufficient statistic—the distribution of bidder valuations—that is estimated in a first stage to solve for the optimal reserve price and test it in a second stage. A qualitative difference from Ostrovsky and Schwarz (2023) is that the space of reserve prices is one-dimensional while the space of disclosure policies is infinite-dimensional, so our sufficient-statistic approach avoids an intractable task of experimenting over a large design space.¹⁰ Our approach also avoids estimating a fully-specified structural model of behavior, a benefit that has been previously recognized in the context of welfare analysis (Chetty, 2009).

2 Conceptual Framework for Human-AI Collaboration

This section develops our conceptual framework for designing human-AI collaboration to solve binary classification and prediction problems, such as classifying a statement as true or false. We take the perspective of a designer who has access to AI predictions and designs a policy to disclose information about these prediction to a human decision-maker, who then makes a classification decision. We also consider settings where the designer has the authority to make the classification directly on the basis of the AI prediction, without involving a human.

⁹Agarwal et al. (2023) finds under-response to AI among professional radiologists, but in contrast to our results they find that radiologists are under-confident in AI information and are not overconfident in their own information. Our approach for estimating these biases also differs in data requirements, discussed below.

¹⁰Dubé and Misra (2023) uses experimental data on a subset of policies—prices—to estimate a function that predicts the outcome of interest—revenue—and tests the optimal policy in a second-stage experiment. This approach is not tractable in our setting because the set of disclosure policies is high-dimensional. Misra and Nair (2011) estimates a structural model of dynamic effort allocation to design an optimal dynamic incentive contract and tests it in a second stage.

The designer’s objective is to maximize the expected accuracy (the probability of correct classification) of the overall human-AI collaborative system.

2.1 A Sufficient Statistic

Each case i in a set I must receive a binary classification $a_i \in \{0, 1\}$ (e.g., False or True). The correct classification (*ground truth, state*) is denoted $\omega_i \in \{0, 1\}$, with prior $\Pr(\omega = 1) = \phi$. An AI tool produces an assessment $\theta_i \in [0, 1]$ of the probability that $\omega_i = 1$. The assessment is calibrated: $\Pr(\omega_i = 1|\theta_i) = \theta_i$. The ground truth ω_i is independent across cases, and the AI assessment θ_i is independent across cases conditional on ω_i . Denote the distribution of each AI assessment θ_i by F . This distribution reflects the quality of the AI’s information about the state. For example, if the AI assessment is always perfectly accurate then $\theta_i \in \{0, 1\}$ with probability 1, while if the AI assessment contains no information then $\theta_i = \phi$ (the ex ante probability that $\omega_i = 1$) with probability 1. In general, a better AI (one that provides more information about ω_i in the sense of Blackwell (1953)) corresponds to a more spread-out distribution F . We suppress the case subscript i for the remainder of the current section.

Given an AI assessment θ , the designer either discloses a signal of the assessment to a human decision-maker or automates the decision by making the classification on its own. Signals can potentially take any form, including quantitative statements like, “The AI assessment is $\theta = 0.7$,” as well as qualitative ones like, “The AI assesses that the statement is likely true.” Formally, the designer chooses an *automation/disclosure policy* $\sigma : \Theta \rightarrow \Delta(\{0, 1\} \cup R)$, where R is an arbitrary set of signal realizations, and σ_θ is the probability that a case with AI assessment θ is either automatically classified as false ($\sigma_\theta(0)$), automatically classified as true ($\sigma_\theta(1)$), or delegated to a human-decision maker who receives signal r from the AI ($\sigma_\theta(r)$), for each possible $r \in R$. The designer’s problem is to design an automation/disclosure policy σ to maximize the probability of correct classification, $\Pr(a = \omega)$.

The optimal design depends on the probability that a human decision-maker correctly classifies a case when they receive any possible signal r . In principle, this probability could depend on a wide range of factors, including the entire posterior distribution $\mu_r \in \Delta([0, 1])$ over the AI assessment θ conditional on receiving signal r under automation/disclosure policy σ , as well as “behavioral” factors such as the language in which signals are expressed. However, we maintain the following assumption:

Assumption 1 *The probability that a human decision-maker correctly classifies a case when they receive a signal r from the AI depends only on the posterior probability over the state, $\Pr(\omega = 1|r) = x$. We denote the probability of correct classification at posterior x by $V(x)$.*

Under Assumption 1, the optimal automation/disclosure policy depends on human behavior only through the function V . The function V is thus the key sufficient statistic that allows us to solve for the optimal policy. Following the information design literature (e.g., Dworzak and Martini (2019)), we refer to $V(x)$ as the designer’s *indirect utility* from inducing posterior belief x .¹¹ Under Assumption 1, the indirect utility function V is “structural,” in that it is defined independently of the AI disclosure policy.

Assumption 1 implies that signals matter only through their probabilistic content and not through the language used to express them. It also implies that there is no benefit to disclosing a non-degenerate probability distribution over AI assessments $\mu \in \Delta([0, 1])$ rather than just the mean assessment $\mathbb{E}^\mu[\theta]$, which equals $\Pr(\omega = 1|\theta \sim \mu)$ by the assumption that the AI assessment θ is calibrated. For example, the probability of correct classification when the AI discloses that $\theta = 0.7$ must be the same as the probability of correct classification when the AI discloses that θ is a 50-50 mixture of 0.5 or 0.9. A signal r can thus be identified with the induced posterior $x = \mathbb{E}^{\mu_r}[\theta]$, and a disclosure policy can be summarized as a distribution G of induced posteriors x . This observation greatly simplifies the formulation of the optimal automation/disclosure design problem. In particular, in our experimental design, a signal from the AI to human participants will take the form of a disclosed mean AI assessment $x = \mathbb{E}^{\mu_x}[\theta]$.

A leading example where Assumption 1 is satisfied is when human decision-makers are Bayesians with correctly specified beliefs and obtain a private signal s of ω that is independent of r conditional on x and ω . This holds because, letting $h(s|x, \omega)$ denote the probability of the human decision-maker’s signal s conditional on any (r, ω) where $\Pr(\omega = 1|r) = x$, we have

$$\frac{\Pr(\omega = 1|s, r)}{\Pr(\omega = 0|s, r)} = \frac{\Pr(\omega = 1|r) \Pr(s|r, \omega = 1)}{\Pr(\omega = 0|r) \Pr(s|r, \omega = 0)} = \frac{x}{1-x} \frac{h(s|x, \omega = 1)}{h(s|x, \omega = 0)}, \quad (1)$$

where $\Pr(s|r, \omega) = h(s|x, \omega)$ by the hypothesis that the distribution of s is measurable with respect to x and ω conditional on r and ω .¹² Note that this example allows the possibility that decision-makers exert costly effort in acquiring information about ω , where their effort choice can depend on posterior $x = \Pr(\omega = 1|r)$ (but does not depend on r conditional on x). In contrast, Assumption 1 is typically violated with conditionally dependent private signals.

¹¹Our approach remains valid if the designer’s indirect utility V differs from the probability of correct classification (e.g., the expected squared loss from a human decision-maker’s probability estimate), so long as Assumption 1 holds with this V replacing the probability of correct classification.

¹²In general, for a correctly-specified Bayesian decision-maker, Assumption 1 holds if and only if the private signal s depends on x linearly conditional on ω , so that the distribution H of s conditional on (x, ω) satisfies $H(s|x, \omega) = (1-x)H(s|0, \omega) + xH(s|1, \omega)$ for all (s, x, ω) .

For example, if the human signal s and the AI signal θ are perfectly correlated, then human classification accuracy following a signal that reveals that $\theta = 0.5$ is 0.5 (as then s also equals 0.5), while human classification accuracy following a signal that reveals that θ is a 50-50 mixture of 0 and 1 is 1 (as now s is either 0 or 1), even though these two signals both result in the same posterior $\Pr(\omega = 1|r) = 0.5$.

Assumption 1 is also satisfied if decision-makers make errors in probabilistic reasoning, but nonetheless their response to signals received from the AI depends only on the posterior x . For example, this holds if decision-makers combine their own (conditionally independent) assessment of the state with the posterior AI assessment via a non-Bayesian procedure such as weighted linear or non-linear averaging. A leading example of such an averaging rule is the belief-updating model in Grether (1980): a Grether agent updates their belief according to (1) with heterogeneous exponential weights on the ratios $x/(1-x)$ and $h(s|x, \omega = 1)/h(s|x, \omega = 0)$, so the resulting posterior belief again depends on r only through x .

Our empirical results will show that the predicted accuracy of policies designed based on Assumption 1 closely matches their realized accuracy, even though we will find that our experimental participants are not correctly-specified Bayesian and their information is not conditionally independent of the AI assessment. This agreement between predicted and realized accuracy based on Assumption 1 is a practical validation of the sufficient-statistic approach.

2.2 The Designer’s Problem

We now show how the optimal automation/disclosure policy can be determined as a function of the indirect utility function $V(x)$. Under Assumption 1, an information disclosure policy can be summarized by the distribution G of induced posteriors x . A key result from the information design literature (Blackwell, 1953; Gentzkow and Kamenica, 2016; Kolotilin, 2018) implies that such a distribution G is attained by some disclosure policy if and only if it is a *mean-preserving contraction* of the distribution F of AI assessments θ . Therefore, the maximum expected accuracy attainable by information disclosure alone (when the designer is not permitted to automate the classification decision) is

$$\max_{G \in MPC(F)} \int_0^1 V(x) dG(x), \tag{2}$$

where $MPC(F)$ denotes the set of all distributions that are mean-preserving contractions of the distribution F of AI assessments. For example, under the *full disclosure* policy, where the AI always discloses its assessment, expected accuracy is given by $\int_0^1 V(x) dF(x)$; while

under the *no disclosure* policy, where the AI reveals no information, expected accuracy is given by $\left(\int_0^1 x dF(x)\right) = V(\phi)$.

Next, to analyze the case where selective automation as a function of x is allowed, define $W(x) = \max\{V(x), 1 - x, x\}$. This is the maximum accuracy that an AI with assessment x can attain by either disclosing this assessment to a human ($V(x)$), classifying the statement as false without human input ($1 - x$), or classifying the statement as true without human input (x). When selective automation is feasible, the maximum expected accuracy attainable by the designer is

$$\max_{G \in MPC(F)} \int_0^1 W(x) dG(x). \quad (3)$$

The optimal policy is then given by (i) garbling the AI assessment so that the distribution of posteriors x is given by the solution G , (ii) disclosing x if $V(x) \geq \max\{1 - x, x\}$, and (iii) automating the decision and classifying the statement as false (resp., true) without human input if $x < \min\{1 - V(x), 0.5\}$ (resp., $x > \max\{V(x), 0.5\}$).

If the human decision-maker is Bayesian with correct beliefs about the joint distribution of s , x , and ω , then $V(x) \geq \max\{1 - x, x\}$, because $\max\{1 - x, x\}$ is the accuracy of a Bayesian decision-maker with no information beyond the AI assessment x . Thus, with a rational decision-maker, $W(x) = V(x)$, and the designer never automates a decision. However, if the human decision-maker is irrational or under-responds to information provided by the AI (consistent with evidence from prior experiments (Benjamin, 2019) and studies on human-AI interaction (Agarwal et al., 2023)), then we may have $V(x) < \max\{1 - x, x\}$ and hence $W(x) > V(x)$ for some values of x , so selective automation may be optimal.

The parameters of the framework are thus the distribution of calibrated AI assessments F and the function $V(x)$ describing human decision accuracy as a function of the disclosed posterior x . In our experiment, the distribution of assessments F is given and known. The experiment estimates the function $V(x)$. Given this function, we can calculate the optimal automation/disclosure policy and the optimal disclosure-only policy as described above.

We will also solve for the optimal *no collaboration policy*, where the AI and the human decision-maker do not communicate. This is the optimal policy with automation but no disclosure: that is, the optimal policy among those that selectively automate cases but do not provide any information about those cases delegated to the human decision-maker. We formulate this problem as choosing a set of assessments $\Theta^{\text{aut}} \subset [0, 1]$, where cases with AI assessments $\theta \in \Theta^{\text{aut}}$ are automated and cases with AI assessments $\theta \notin \Theta^{\text{aut}}$ are delegated to the human-decision maker, who is informed only of the posterior among delegated cases,

$\mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}]$. The set Θ^{aut} is chosen to solve

$$\max_{\Theta^{\text{aut}} \subset [0,1]} \mathbb{E}[\max\{\theta, 1 - \theta\} | \theta \in \Theta^{\text{aut}}] \Pr(\theta \in \Theta^{\text{aut}}) + V(\mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}]) \Pr(\theta \notin \Theta^{\text{aut}}).^{13} \quad (4)$$

This sufficient-statistic approach differs in two ways from the existing literature, which studies policies that selectively automate cases as a function of the AI assessment (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023). First, our approach accounts for how human decision-makers’ beliefs respond to the designer’s automation/disclosure policy. For instance, in equation (4), human accuracy on delegated cases equals $V(\mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}])$, which depends on the set of automated cases Θ^{aut} . Appendix D.1 discusses how this response was neglected in previous work and quantifies the implications of taking it into account. Second, unlike previous approaches, we do not need to collect data under multiple disclosure policies to find the optimal policy. Instead, the estimated function V based on data under full disclosure and the distribution F are used to predict accuracy for any counterfactual disclosure or automation policy.

2.3 Discussion of the Optimal Design

We now describe how the shape of the function V determines the optimal automation/disclosure policy and preview our empirical results on the shape of V .

First, full disclosure without automation is optimal if and only if V is convex and $V(x) \geq \max\{1 - x, x\}$ for all x . For example, these conditions hold if the human decision-maker is Bayesian and the distribution of her private signal s is independent of θ conditional on ω .¹⁴

Second, if V is convex but $V(x) < \max\{1 - x, x\}$ for some x , then a mix of full disclosure and automation is optimal: the designer should disclose assessments θ where $V(\theta) \geq \max\{1 - \theta, \theta\}$ and should automate the decision if $V(\theta) < \max\{1 - \theta, \theta\}$. For example, this case can arise if human decision-makers observe conditionally independent private signals but under-respond to AI-provided information. Figure 1a and 1b illustrate some functions V where full disclosure without automation and with automation are optimal.

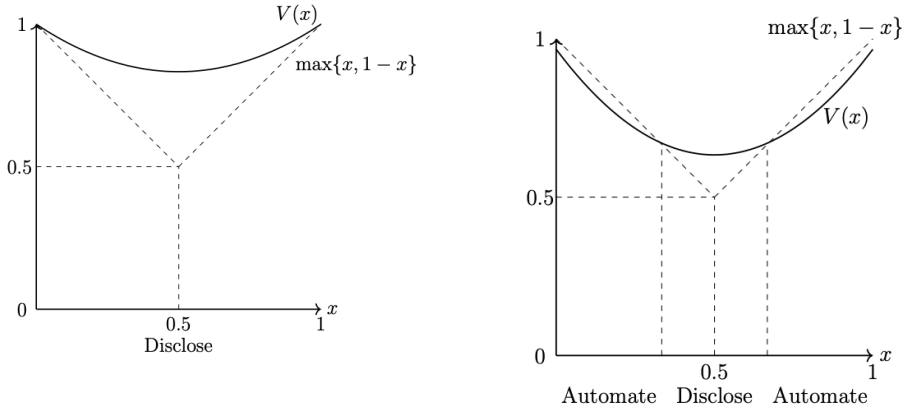
To preview, our empirical results will indicate that in our setting V is (approximately) convex, so fully disclosing the AI assessment is optimal. We also find that there are values

¹³This formulation assumes that the designer does not randomize cases with any assessment θ between automation and delegation to humans. In our setting, the gains from such randomization are trivial.

¹⁴Intuitively, V is convex because a Bayesian cannot do better by ignoring any AI information, and $V(x) \geq \max\{1 - x, x\}$ for all x because a Bayesian cannot do better by ignoring her own information. Conversely, any convex function V satisfying $V(x) \geq \max\{1 - x, x\}$ for all x is the probability of correct classification for some conditionally independent distribution for s (Kolotilin et al., 2017).

Figure 1: Indirect Utilities where Full Disclosure with and without Automation is Optimal

(a) Full Disclosure with No Automation (b) Full Disclosure with Automation



Note: In Panel (a), full disclosure with no automation is optimal because V is convex and $V(x) \geq \max\{1-x, x\}$ for all x . In Panel (b), full disclosure is optimal for AI assessments x where $V(x) \geq \max\{1-x, x\}$, and automation is optimal for AI assessments x where $V(x) < \max\{1-x, x\}$. The function V we estimate is qualitatively similar to the one in Panel (b).

of x where $V(x) < \max\{1-x, x\}$, so automation is valuable. Qualitatively, the function V that we estimate has a similar shape as the function V in Figure 1b.

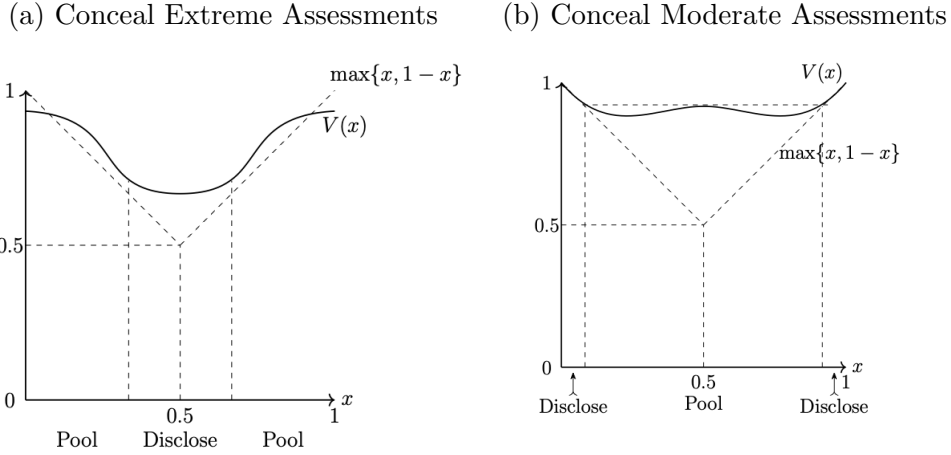
More generally, other disclosure policies can be optimal. In particular, the function V may be non-convex if human effort is sufficiently sensitive to x or if humans exhibit more complex behavioral biases. Figures 2a and 2b illustrate such functions V . If V is non-convex then full disclosure is suboptimal, so optimal information disclosure takes a more complex form. For example, Kolotilin (2018) characterizes when it is optimal to pool extreme states and disclose intermediate states, or vice versa.¹⁵ Empirically, Dell’Acqua (2022) finds a setting where human effort is sufficiently sensitive to the disclosed AI signal that overall accuracy is higher with a less precise AI signal, which would imply that V is non-convex under Assumption 1, and Agarwal et al. (2023) likewise finds that withholding the AI signal improves accuracy for some cases.

Finally, while we focus on binary classification problems, a similar approach applies for multi-class classification. In the general multi-class case with n possible classifications, the state ω lies in an arbitrary finite set Ω with n element, and the designer’s problem becomes a general Bayesian persuasion problem as formulated in Kamenica and Gentzkow (2011).¹⁶

¹⁵In these settings, it can be optimal for the designer to delegate some cases with assessments θ where $V(\theta) < \max\{1-\theta, \theta\}$ to a human—even though the AI’s accuracy on these cases would be higher if it automated them—because pooling these cases with other cases where $V(\theta) > \max\{1-\theta, \theta\}$ increases accuracy on the latter cases, which more than compensates for the reduced accuracy on the former cases.

¹⁶Here, the designer’s indirect utility function is a function $V : \Delta(\Omega) \rightarrow \mathbb{R}$ defined on the $n-1$ -dimensional simplex (e.g., the probability of correct classification when AI assessment $\mu \in \Delta(\Omega)$ is disclosed to the decision-maker), and the designer’s problem is to maximize $\int_{\mu \in \Delta(\Omega)} V(\mu) d\tau(\mu)$ over disclosure policies $\tau \in \Delta(\Delta(\Omega))$,

Figure 2: Indirect Utilities where Partial Disclosure is Optimal



Note: In Panel (a), it is optimal to disclose moderate assessments and separately pool extreme low and high assessments. This pattern can arise if AI under-response is more extreme at extreme AI assessments. In Panel (b), it is optimal to disclose extreme assessments and pool moderate assessments. This pattern can arise if AI information strongly crowds out human effort.

The main difference is that the indirect utility function to be estimated and the set of possible disclosure policies to be optimized over are both lower-dimensional in the binary case.

3 Experimental Design

We design a two-stage experiment to implement and validate the sufficient-statistic approach in the context of human-AI collaboration in fact-checking. Stage 1 estimates the function $V(x)$ —the probability of correct classification as a function of the disclosed mean AI assessment $x \in [0, 1]$. Stage 2 then tests the automation/disclosure policies that we find to be optimal under the $V(x)$ function estimated in Stage 1, as well as some benchmark policies.

The two stages are nearly identical except for the AI assistance provided to participants. In Stage 1, the AI assessment θ is always disclosed to participants: in other words, the automation/disclosure policy is Full Disclosure + No Automation. In Stage 2, we test the five automation/disclosure policies mentioned in the introduction: Full Disclosure + Automation (the predicted optimal policy with automation), No Disclosure + Automation, Full Disclosure + No Automation (the predicted optimal policy without automation), No Disclosure + No Automation, and Stoplight.

We pre-registered this design and updated the plan to describe the specific policies tested subject to the Bayes plausibility constraint $\int_{\mu \in \Delta(\Omega)} \mu d\tau(\mu) = \phi$, where $\phi \in \Delta(\Delta(\Omega))$ is the population distribution of the AI assessment $\mu \in \Delta(\Omega)$.

in Stage 2 as a result of the Stage 1 estimates.¹⁷ The experiment was implemented on Prolific (www.prolific.com) using an interface designed on the o-tree framework (Chen et al., 2016) that can be accessed through a browser. We next describe the interface and the data that we collected during the experiment.

3.1 The Task

In our experiment, participants assess the probability that statements are True or False. For each statement, the participant encounters a screen that includes the statement, an AI assessment of the probability that the statement is True, a link to a Google search for the subject of the statement, and a slider where the participant enters their assessment. Figure 3 presents a screenshot of the experimental interface. For each statement, we record the participant’s assessment $p \in [0, 1]$ and a binary classification $a \in \{0, 1\}$, where $a = 1 [p > 0.5]$.¹⁸

After entering their assessment, participants self-report if they used an external source, including the Google link (Figure 3b). Participants then encounter a feedback screen that includes the AI assessment, the participant’s assessment and classification, and the ground truth state, True or False (Figure 3c).

In addition to assessments and classifications, we also collect three measures of effort: the time taken on each statement, whether the participant clicked the Google search link, and the participant’s self-report of whether they used an external source.

In Stage 1, each participant assesses 30 random statements from our statement database (described in Section 3.3). In Stage 2, each participant assess 40 random statements: eight different statements under each of five different automation/disclosure policies. To economize on statistical power, our design includes within-participant comparisons. The order of the policies is randomized to ensure that our estimated treatment effects are not confounded with learning or fatigue and to preserve a robustness check using a pure across-participant comparison based on the first treatment.

¹⁷The pre-registration can be found at <https://doi.org/10.1257/rct.13990-1.1>. We also pre-registered that we would update the plan after Stage 1 with the Stage 2 policies we test. The updated pre-registration changed the structure of the second stage to test 5 policies rather than the 4 we initially intended to test. We also reduced the number of statements per policy to 8 rather than 10 to maintain the overall duration of the experiment for each participant. Unless otherwise noted, all analyses we present are pre-registered.


¹⁸Participants enter their probability assessment through the slider in Figure 3a. The slider button and the text below the slider (“Likelihood true: $p\%$ ” and “Your classification: a ”) appear after the participant clicks on the slider to avoid priming.

Figure 3: Screenshots of Experimental Interface

(a) Assessment Screen

Statement 6/45

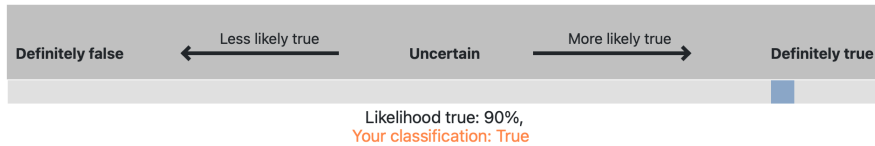
French-Canadian musician Marc Remillard began producing music in 2007, when he released his first track, A Little Less Glitches and has since released numerous singles, EPs and remixes.

AI assessment: Likelihood statement is true is 65%  ⓘ

Link to google search for "French-Canadian musician Marc Remillard":

Google Search

Your assessment:



Submit

(b) Self Reported Effort Screen

Statement 6/45

French-Canadian musician Marc Remillard began producing music in 2007, when he released his first track, A Little Less Glitches and has since released numerous singles, EPs and remixes.

AI assessment: Likelihood statement is true is 65%  ⓘ

Did you use any external sources (including the Google link) to check this statement? Reminder: external sources are allowed and your response to this question does not affect your payment.

No, I did not use an external source

Yes, I used an external source

(c) Feedback Screen

Statement 6/45

French-Canadian musician Marc Remillard began producing music in 2007, when he released his first track, A Little Less Glitches and has since released numerous singles, EPs and remixes.

AI assessment: 65%

Your assessment: 90%

Your classification: True

✗ Incorrect! Your classification was True and the correct answer was False

Next

3.2 Participant Recruitment, Training, and Incentives

We recruit participants from the Prolific platform. We use Prolific’s filters to ensure that each person participates at most once and no one participates in both the first and second stages of the experiment. We recruit a sample representative of the United States adult population on the dimensions of sex, age, and ethnicity.¹⁹ A summary of demographic information of the study participants is presented in Appendix Table A.2. We saw minimal attrition, with 97.7% of participants who grant consent and began the study completing Stage 1 and 95.8% completing Stage 2.

At the beginning of the experiment, participants receive an overview of the task and the compensation rule. They are then provided with additional information about the task, the interface, and the prior distribution of true and false statements in the database. We next introduce the AI fact checker, explaining that it provides a calibrated assessment of the likelihood that a statement is true. In Stage 2, we explain that participants will encounter multiple AI fact-checkers (see Appendix F.4). Next, we explain the compensation rule in broad terms and highlight that the expected payment increases with the accuracy of the assessment. We also provide a button that opens a window that provides full details of the compensation rule. We then test the participant’s understanding of the task and the AI fact-checker through a series of comprehension questions. These questions test if participants understand that the AI is calibrated, that they can use outside resources, and that they understand the compensation rule. Finally, before beginning the experiment, each participant assesses five practice statements to ensure familiarity with the experimental interface. The full experimental instructions are presented in Appendix F.

Participants are incentivized in two ways to exert effort and provide accurate assessments. The first is a bonus of 35 cents for each correctly classified statement. The second is a lottery for an additional \$20, where the probability of winning the lottery depends on the accuracy of the participant’s probability assessments following Hossain and Okui (2013).

3.3 The Statements

We use the set of statements collected and labeled in the FEVEROUS database (Aly et al., 2021). The data contain approximately 80,000 statements that are constructed by asking annotators to generate statements from a snippet of highlighted Wikipedia text or tables. A separate set of annotators are asked to label each statement as either Supported (True),

¹⁹Certain segments are under-represented on Prolific, including older adults. We maintained the representative target until 95% of slots were filled. We filled the remaining slots in both stages with non-representative participants.

Refuted (False), or Not Enough Information (NEI).²⁰

Aly et al. (2021) describes the extensive quality controls taken to ensure high-quality statements and labels (Aly et al., 2021). In addition, we remove statements that are not suitable for our study. We first remove the approximately 3% of statements with an NEI label. We also remove statements with any spelling or grammatical errors flagged by either the rules-based LanguageTool API or GPT-4o.²¹ Finally, we remove statements that we determine to be of poor quality, which are mostly statements where the ground truth can change over time, such as statements that reference an individual’s age. In the final database of statements from which we sample, 65.4% of statements are True.²²

3.4 The AI Fact-Checker

We use OpenAI’s GPT-4o as our AI fact-checker because it generated more accurate assessments than other alternatives, including the fact-checker in Aly et al. (2021). For each statement, we queried the OpenAI API with the prompt, “True or False: [statement]” and store the top 20 most likely next tokens along with the probability of each token. We calculate a raw score θ_i^r for each statement i as

$$\theta_i^r = \frac{\sum_j p_{ij} 1[\text{token}_{ij} = \text{true}]}{\sum_j p_{ij} 1[\text{token}_{ij} \in \{\text{true}, \text{false}\}]},$$

where token_{ij} is the j^{th} most likely next token, and p_{ij} is the probability GPT-4o assigns to the j^{th} token.²³ We then calibrate θ_i^r by binning it into 200 bins and calculating the share of statements in each bin that are true to yield the calibrated AI assessment θ_i . Figure 4 shows the distribution of θ_i .

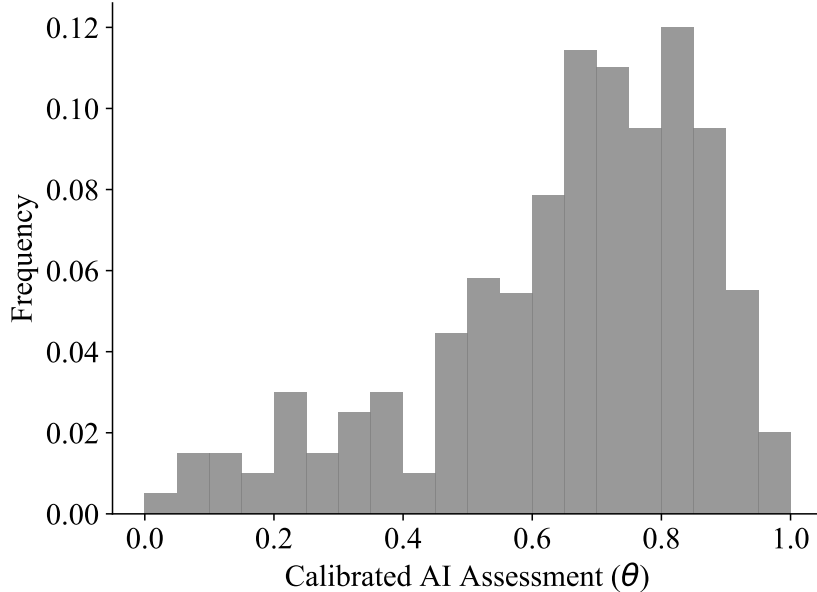
²⁰Supported statements require all information within the statement to be verified and supported by evidence. Refuted statements require only a single piece of information within the statement to be refuted by evidence. Statements where not enough information is available on Wikipedia to label the statement as either True or False are labeled Not Enough Information.

²¹We queried GPT-4o with the prompt “True or False. The following statement has no grammatical or spelling errors:” followed by each statement. We discarded statements that GPT-4o assessed to be more likely than not to contain a spelling or grammatical error.

²²Our independent review of 50 randomly drawn statements, half of which are true, found three cases in which our assessed label differed from the label in FEVEROUS and three cases where there was not enough information or ambiguous wording.

²³GPT-4o is highly likely to suggest tokens in the set $\{\text{true}, \text{false}\}$. In our sample, $\sum_j p_j 1[\text{token}_j \in \{\text{true}, \text{false}\}]$ is greater than 0.9 for 99.5% of statements and greater than 0.99 for 94.7% of statements.

Figure 4: Distribution of Calibrated AI Assessments



Note: Histogram of calibrated AI assessments (from GPT-4o) for the final population of statements in our database.

4 Stage 1 Results

In Stage 1, we estimate the function V introduced in Section 2.1 and calculate optimal and approximately optimal information disclosure policies with and without automation. We also calculate the predicted treatment effect of each policy and document how our effort measures respond to the AI assessment θ (which equals x under full disclosure).

4.1 Overall Accuracy and Effort

Table 1 describes participants’ accuracy and effort. Participants correctly classified 73.5% of statements. This overall accuracy is similar to the accuracy of 73.3% that would result if participants simply repeated the AI assessment. (Recall that the AI assessment is fully disclosed to participants in Stage 1.) However, this similarity masks large heterogeneity in accuracy by AI assessment x —that is, large variation in $V(x)$ over x —which is key for determining the optimal automation/disclosure policy. We discuss the shape of V in the next subsection.

Participants classified 69.6% of cases as True. This exceeds the share of true cases in the database, 65.4%, which was conveyed to participants. The mean participant assessment of 63.0% is closer to the share of true cases.

Participants appear to exert considerable effort: they reported using external information sources in 63.7% of cases; clicked the provided Google search link in 36.0% of cases; and took an average of 46.8 seconds fact-checking each statement.²⁴

Table 1: Stage 1 Summary Statistics

	Stage 1	
	Mean	SD
	(1)	(2)
Correct Classification	0.735	0.441
Classified as True	0.696	0.460
Assessment	0.630	0.329
Used External Sources	0.637	0.481
Clicked Google Link	0.360	0.480
Time Taken (s)	46.791	43.959
Observations	45030	
Participants	1501	
Cases per Participant	30	

Note: Summary statistics of the Stage 1 data. Correct Classification is an indicator for whether the classification matches the ground truth. Classified as True is an indicator for whether the probability reported exceeds 0.5. Assessment is the reported probability true. Used External Sources is an indicator for whether the participant self-reported using external sources. Clicked Google Link is an indicator for whether the participant clicked the provided Google link. Time taken (s) for a statement is measured in seconds and winsorized at the 5th and 95th percentiles.

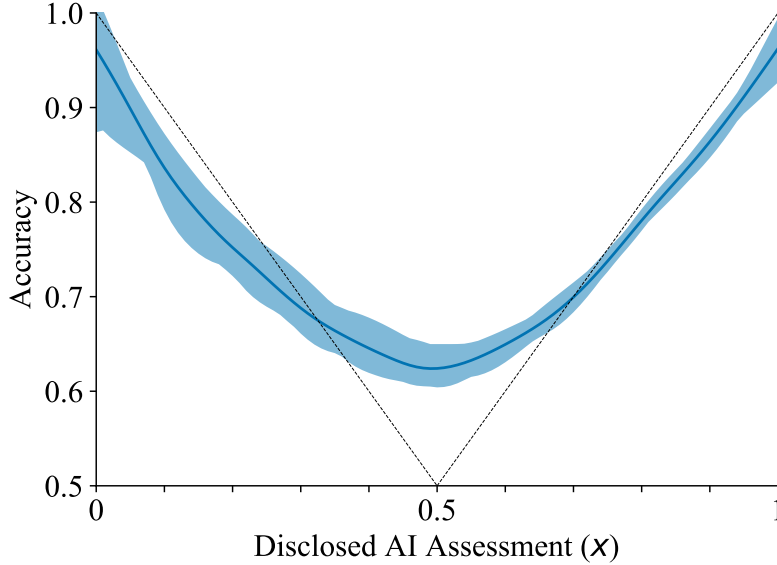
4.2 Accuracy and Effort by AI Assessment x

Figure 5 presents our estimate of the sufficient statistic V , obtained using a local linear regression. The estimated function \hat{V} has a qualitatively similar shape as the function V in Figure 1b. There are two important features. First, \hat{V} is approximately convex, and a statistical test does not reject that V is convex ($p=0.5$).²⁵ Recall that if V is convex then

²⁴The median participant in Stage 1 took 44 minutes to finish the experiment, including training, comprehension questions, and the 5 practice statements.

²⁵We tested convexity by comparing the objective function from estimating $V(x)$ with and without the convexity constraint. Specifically, we estimated $V(x)$ using local linear regression subject to a global convexity constraint via a quadratic programming problem: $\min_{g,\beta} \sum_i (y_i - g - \beta(x_i - t))^2 K_h(x_i - t)$ subject to $(g_{j+1} - 2g_j + g_{j-1}) / (t_{j+1} - t_j)^2 \geq 0$ for all j , where $K_h(\cdot)$ is a Gaussian kernel with bandwidth h . The test statistic (the value of the objective function) is compared to a null distribution generated using a bootstrap distribution of the objective function without the convexity constraint. The null distribution and test statistic are plotted in Figure B.4.

Figure 5: First Stage Estimate of V



Note: V is estimated using local linear regression from Stage 1 data. The bandwidth is chosen via leave-one-out cross validation to minimize mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level (Montiel Olea and Plagborg-Møller, 2019). The dashed lines indicate the accuracy of $\max\{x, 1-x\}$ that would result under automation.

full disclosure of the AI assessment is optimal for all non-automated cases.

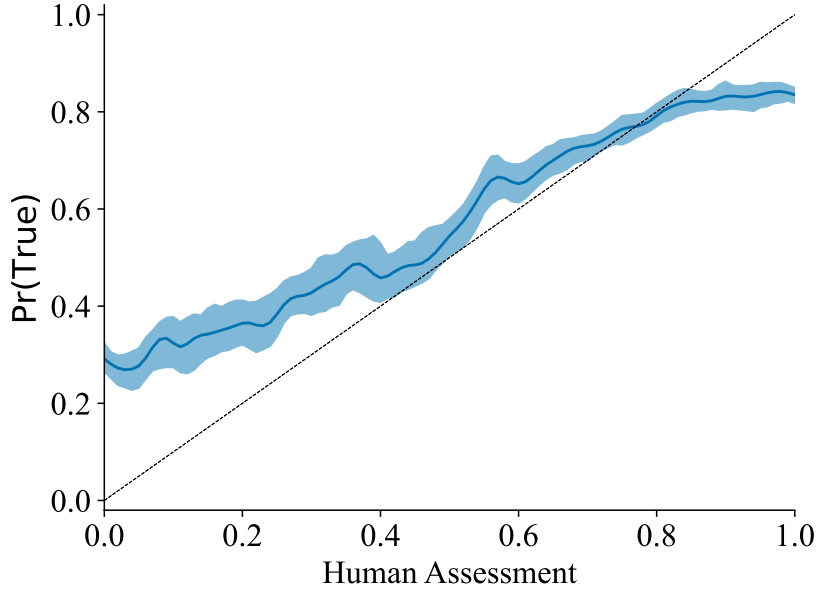
We thus obtain a key implication for optimal design: in any optimal automation/disclosure policy, the AI assessment of any non-automated case should be fully disclosed to the human decision-maker.²⁶

Second, on cases where the AI is confident, participants with AI assistance perform significantly worse than they would if they just followed the AI. Figure 5 shows that $V(x) < \max\{x, 1-x\}$ whenever $x < 0.33$ or $x > 0.69$. Automation would improve accuracy on these cases. At the same time, participants significantly outperform the AI on cases where the AI is uncertain: for example, $V(0.5) = 0.62$, which substantially exceeds the accuracy of 0.5 that would result from automating these cases.

The fact that participants would do better by just following the AI for some range of AI assessments implies a degree of under-response to the AI. This finding echoes under-response to information in experiments on belief updating (Benjamin, 2019) and automation neglect in experiments involving predictive AI assistance (e.g. Agarwal et al., 2023).

²⁶Appendix A.2 contains an estimate of V when the designer’s objective is to minimize the deviation of the probability assessment from the ground truth (i.e., $V(x) = E[|p_{ij} - \omega_i| \mid x]$). We also find V to be convex for this alternative objective.

Figure 6: Calibration Curve of Human Assessments



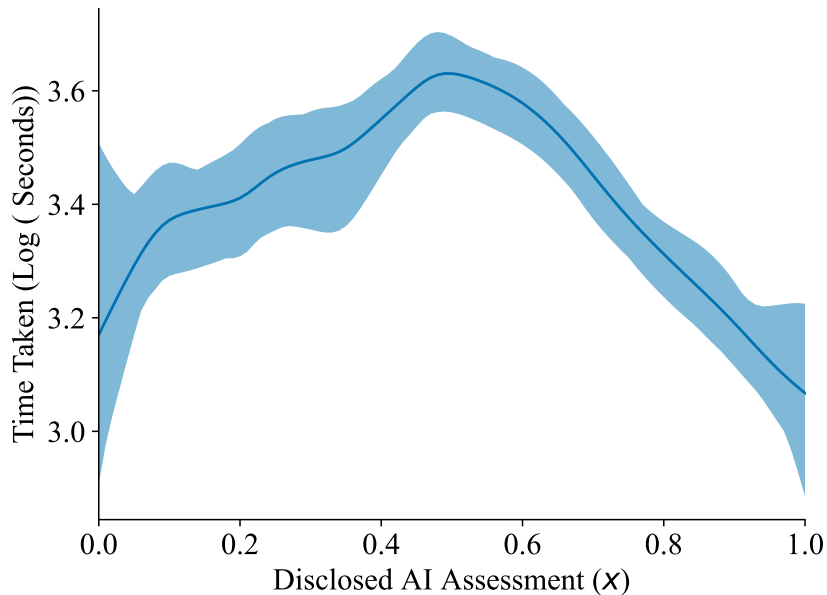
Note: Calibration curve in Stage 1. Local-linear regression of ω_i on reported assessments using a Gaussian kernel. Bandwidth is selected to minimize cross-validated mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level.

While Figure 5 shows that participants under-respond to the AI, it does not indicate whether this occurs because participants *under-weight* the AI’s information—consistent with *automation neglect* as found in Agarwal et al. (2023)—or because participants *over-weight* their own information—consistent with the version of overconfidence known as *over-precision* in the behavioral economics literature (Moore and Healy, 2008). In particular, the function V in Figure 5 can be generated by either a quasi-Bayesian with correct beliefs about the precision of their own signal but erroneously low beliefs about the precision of the AI signal, or a quasi-Bayesian with correct beliefs about the precision of the AI signal but erroneously high beliefs about the precision of their own signal.

Examining the participants’ reported assessments suggests overconfidence. Figure 6 plots the calibration curve (true probability against reported probability) for Stage 1 participants. The slope of the calibration curve is less than 1, indicating overconfidence. For example, 29% of statements that participants report are definitely false (reported $p = 0$) are actually True, and 16% of statements that participants report are definitely True (reported $p = 1$) are actually False. While Figure 6 suggests overconfidence, it does not speak to automation neglect, and it does not quantify the extent of overconfidence. We address these questions using a structured model of belief updating in Section 6.

We also find evidence of effort crowd-out as the AI assessment x moves away from 0.5, the point of maximum uncertainty. Figure 7 shows the effect on time taken is 10-15 percentage points lower when $x = 1$ as compared to $x = 0.5$. This effect is similar for our other measures (see Figures A.1a and A.1b). This reduction in effort for confident AI assessments is another reason why automation outperforms human-AI collaboration. In Section 6.4, we estimate the effect of disclosing the AI assessment on the precision of participants’ private signals via the induced reduction in participant information-acquisition effort.

Figure 7: Effort Response – Time Taken by AI Assessment



Note: Log time taken (in seconds) to assess a statement by x in Stage 1, estimated via local linear regression. The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

4.3 Optimal and Simple Policies

We now use the estimate of V from Figure 5 to solve for the optimal policies, both when automation is feasible and when it is infeasible: that is, we solve the problems (2) and (3) for the estimated function V . Since the estimated function V is convex, optimal policies fully disclose the AI assessment of any non-automated case to the human decision-maker.

We compare these optimal policies with the optimal no-collaboration policies where the AI discloses no information to the human decision-maker: that is, the No Disclosure + No Automation policy and the No Disclosure + Full Automation policy that solves problem (4).

Finally, we also consider the Stoplight policy where the AI can only disclose one of three

signals. In total, we consider the four policies illustrated in Figure 8, as well as Full Disclosure + No Automation (the optimal policy without automation), which we also ran in Stage 1.

The first two policies allow automation. Here we compare the optimal policy (Full Disclosure + Automation) and the optimal no-collaboration policy (No Disclosure + Automation).

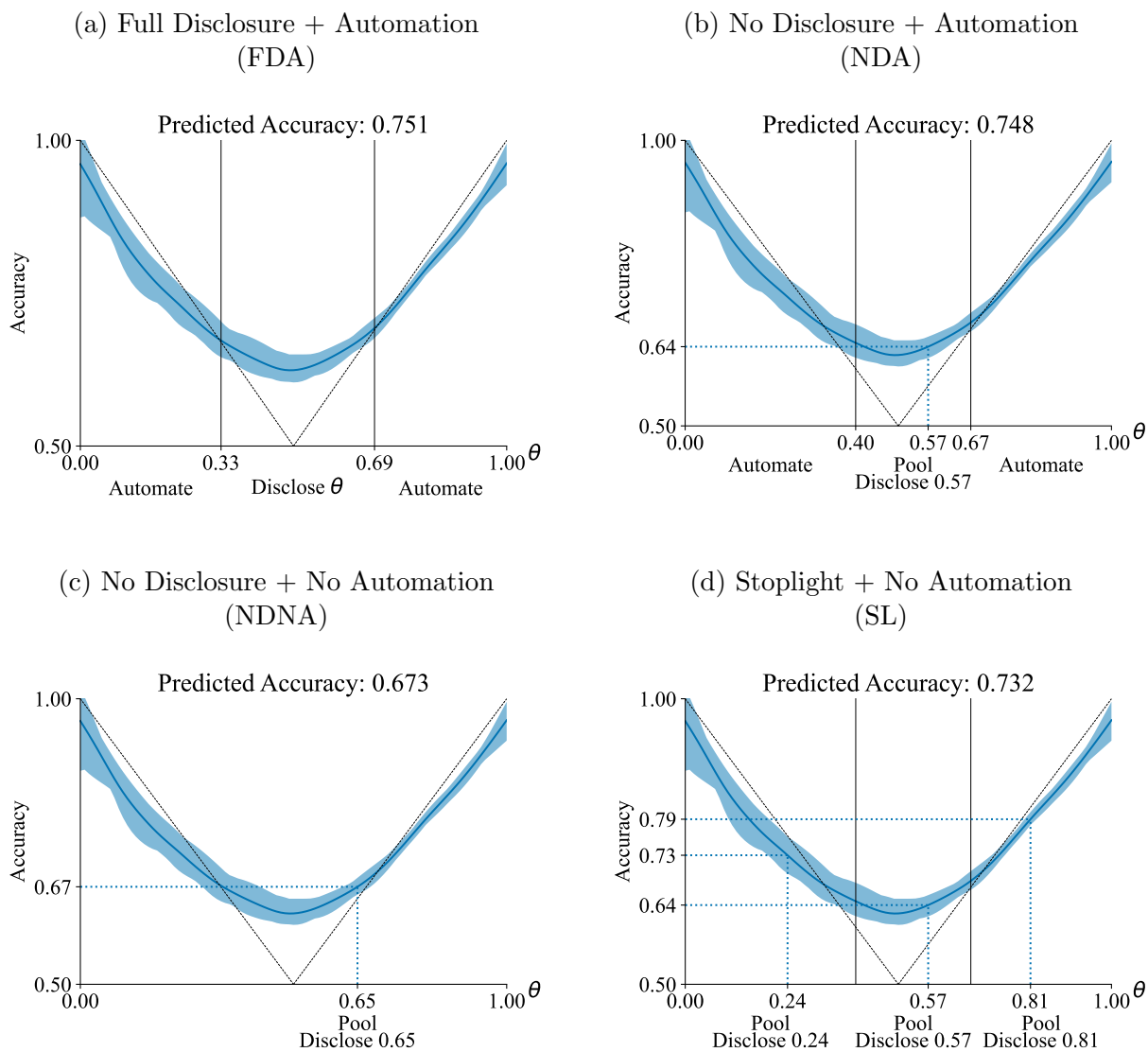
- (a) **Full Disclosure + Automation (FDA)**: The optimal policy (i.e., the solution to (3)) discloses θ if $V(\theta) > \max\{\theta, 1 - \theta\}$ —which we find holds if $\theta \in [0.33, 0.69]$ —and automates the case otherwise. The predicted accuracy of this policy is 75.1%.
- (b) **No Disclosure + Automation (NDA)**: The optimal no-collaboration policy solves problem (4). We find that the optimal set of cases to automate is $\Theta^{\text{aut}} = [0, 0.39] \cup [0.68, 1]$. The mean AI assessment conditional on $\theta \notin \Theta^{\text{aut}}$ is 0.57. The predicted accuracy under this policy is 74.8%. Since this is only 0.3 percentage points lower than the optimal policy of FDA, the predicted value of direct human-AI collaboration is very small.

The intuition for why predicted accuracy under FDA or NDA is almost identical is that the estimated function V is relatively flat on the intervals of non-automated cases, $[0.33, 0.69]$ (for FDA) or $[0.39, 0.68]$ (for NDA). Since the benefit of disclosing information comes from the convexity of V , this implies that the benefit of disclosing AI assessments on the interval of non-automated cases is small.

We highlight that it is optimal to automate cases with a wider range of AI assessments under NDA than under FDA. The key reason is that a marginal case θ at the boundary of the automation region under full disclosure is correctly classified with probability $V(\theta) = \max\{\theta, 1 - \theta\}$, while if this case were delegated to a human under no-disclosure it would be correctly classified with probability only $V(\mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}])$, which is less than $V(\theta)$ for a marginal case θ . So, automating such cases is strictly better under no disclosure. In addition, the decision to automate or delegate marginal cases affects $\mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}]$. Since $V(x)$ is positively sloped at $x = \mathbb{E}[\theta | \theta \notin \Theta^{\text{aut}}]$, this effect favors automating more marginal low- θ cases and fewer marginal high- θ cases under no disclosure, which explains why the lower threshold of the automation region increases substantially—from 0.33 to 0.39—as we move from FDA to NDA, while the upper threshold of the automation region only slightly decreases from 0.69 to 0.68.

The remaining policies consider the case where automation is infeasible. Here we consider the optimal policy (Full Disclosure + No Automation), the no-collaboration policy (No Disclosure + No Automation), and a simple policy that approximates the optimum (Stoplight).

Figure 8: Stage 2 Experiment Overview



Note: Figures summarize the set of policies considered in the experiment. The function V is estimated using local linear regression from Stage 1 data. The bandwidth is chosen via leave-one-out cross validation such that the mean squared error is minimized. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level. The dashed lines indicate the accuracy under automation of $\max\{\theta, 1 - \theta\}$. The dotted lines indicate the assessments disclosed to participants and the associated accuracy predicted by V .

- (c) **Full Disclosure + No Automation (FDNA)**: This is the optimal policy without automation (i.e., the solution to (2)), which is the policy used in Stage 1. The predicted accuracy of this policy equals the average accuracy in Stage 1, 73.5%.
- (d) **No Disclosure + No Automation (NDNA)**: With no disclosure or automation, participants are only informed of the share of True cases in the database, which is 65.4%. The predicted accuracy of this policy is 67.3%.
- (e) **Stoplight (SL)**: The final policy we consider illustrates the feasibility of approximating full disclosure using a very simple signal distribution. Specifically, we calculate the optimal partition of AI assessments into K intervals and disclose the average AI assessment within each interval. The resulting accuracy is

$$\max_{\{\theta_k\}_{k=0}^K: \theta_0=0, \theta_K=1} \sum_{k=1}^K \Pr(\theta \in [\theta_{k-1}, \theta_k]) V(\mathbb{E}[\theta | \theta \in [\theta_{k-1}, \theta_k]]).$$

Note that $K = 1$ gives NDNA, while $K = \infty$ gives FDNA.

We consider “Stoplight” with $K = 3$ for two reasons. First, predicted accuracy with $K = 3$ is 73.2%, which we expect to be indistinguishable from the predicted accuracy of 73.5% when $K = \infty$.²⁷ Intuitively, since the estimated function $V(\theta)$ is well-approximated by a piecewise linear function with three “pieces,” disclosing only which piece contains the AI assessment θ is an approximately optimal policy. Second, Stoplight can be interpreted as a system in which the AI reports only that each case is either “Likely False,” “Uncertain,” or “Likely True” (or “Red,” “Yellow,” or “Green”), which resembles some collaborative systems used in practice.²⁸ The optimal Stoplight policy partitions the AI assessment into the intervals $[0, 0.40)$, $[0.40, 0.68)$ and $[0.68, 1.00]$, with mean assessments 0.24, 0.57, and 0.81, respectively.²⁹

We emphasize five qualitative predictions for the design of human-AI collaboration:

1. **Automation is valuable.** Predicted accuracy under the optimal policy with automation (75.1% under FDA) significantly exceeds that under the optimal policy without automation (73.5% under FDNA).

²⁷Predicted accuracy for other values of K are shown in Figure B.5.

²⁸For example, several pre-trial risk assessment tools report risk levels in coarse bins, including the Pre-Trial Risk Assessment (Lowenkamp, 2009) and the Public Safety Assessment’s Release Conditions Matrix (Policy and Research, 2020)

²⁹It is a coincidence that the middle interval under Stoplight coincides with Θ^{aut} under No Disclosure + Automation.

2. **Human information is valuable.** Predicted accuracy under the optimal policy with automation (75.1%) significantly exceeds that achievable with AI alone (73.3%).
3. **Human-AI collaboration does not outperform selective automation.** Predicted accuracy under the optimal policy with automation (75.1%) does not significantly exceed that under the optimal no-collaboration policy (74.8% under NDA).
4. **AI assistance is valuable when automation is infeasible.** Predicted accuracy under the optimal policy without automation (73.5%) significantly exceeds that without AI assistance (67.3% under NDNA).
5. **Simple disclosure policies are approximately optimal.** The predicted accuracy under the optimal policy without automation (73.5%) does not significantly exceed that under SL (73.2%).

In addition, from the perspective of validating Assumption 1, it is worth highlighting that the quantitative predictions from the above policies are all out-of-sample (except for FDNA). In particular, the no-disclosure and Stoplight policies provide counterfactual AI assessments to our participants. The accuracy predictions under these policies are thus particularly demanding tests of our framework.

4.4 Restrictions on the Design Space

Three restrictions on our design space merit discussion. First, while we let the AI flexibly disclose information to human decision-makers, we do not consider systems that elicit humans’ assessments and combine them with the AI’s information. That is, we consider “one-way” communication from AI to humans, not “two-way” communication. However, in Section 5.2, we consider the maximum accuracy attainable with access to both human and AI assessments under FDNA, and show that this accuracy is indistinguishable from that under FDA (the optimal policy without elicitation). Thus, in our setting, one-way communication turns out to be without loss of optimality.

Second, we restrict to disclosure policies where the AI assessment is calibrated. Appendix D.2 analyzes policies where the designer can exaggerate the AI assessment to offset the under-response to AI information documented above. However, the benefit of exaggeration will wear off over time if humans learn that AI assessments are not calibrated.

Third, we do not tailor the policy to predictable heterogeneity across participants. Figure C.9 shows that accuracy and the sufficient statistic V are predictable as functions of baseline comprehension questions, effort, or accuracy on initial statements. However, Table C.12

shows that policies that are tailored to this heterogeneity yield predicted accuracy similar to that of the pooled policies we consider.³⁰

5 Stage 2 Results

In Stage 2, we test each of the above policies—FDA, NDA, FDNA, NDNA, and SL. Our goals are (i) to compare their accuracy to the predictions based on Stage 1 data described in Section 4.3, (ii) to compare them to a benchmark of the potential gains from optimally combining human and AI signals, and (iii) to document the effects of these policies on effort.

We estimate the average outcome for each policy $k \in \{\text{FDNA, FDA, NDA, NDNA, SL}\}$ in Stage 2 using the regression:

$$y_{ij} = \sum_{k \in \{\text{FDNA, FDA, NDA, NDNA, SL}\}} 1[\text{policy}(i, j) = k] \gamma_j + \varepsilon_{ij}, \quad (5)$$

where y_{ij} is an outcome for statement i by participant j , β_k is the average outcome under policy k . We cluster standard errors to allow for $\text{Cov}(\varepsilon_{ij}, \varepsilon_{i'j'}) \neq 0$ if either $i' = i$ or $j' = j$, but set it to zero otherwise.³¹ Estimated treatment effects relative to FDNA are therefore given by $\beta_k - \beta_{k_0}$, where k_0 is the baseline FDNA policy.

5.1 Validity of the Sufficient-Statistic Approach

Table 2 presents the estimated accuracy under each of the five automation/disclosure policies tested in Stage 2 (column 1) and compares them to the predictions based on the function V estimated using either FDNA in Stage 2 (column 2) or FDNA in Stage 1 (column 4). The p-values for a test of the differences between the experimental estimates and each of the two predictions are shown in columns 3 and 5.

The experiment confirms all of our qualitative predictions:

1. **Automation is valuable.** Accuracy under FDA significantly exceeds that under FDNA. The estimated difference is 2.7 percentage points ($p < 0.001$). The predicted

³⁰The envelope theorem provides a rationale for this result. Since the pooled policies are optimized to the full population of participants, the impact of re-optimizing the policy to fit changes in the sufficient statistic is second-order.

³¹For treatment arms involving automation, participants only assess cases that are not automated. The dependent variables of interest are system accuracy and effort. For automated arms, we use the modified outcomes $y_{ij} \Pr(\theta \notin \Theta^{\text{Aut}}) + \bar{y} \Pr(\theta \in \Theta^{\text{Aut}})$, where Θ^{Aut} is the set of automated AI assessments under a given policy, and \bar{y} is the average outcome among automated cases. For accuracy, $\bar{y} = E[\max\{\theta, 1 - \theta\} | \theta \in \Theta^{\text{Aut}}]$; for effort measures, $\bar{y} = 0$.

difference is 2.5 percentage points using Stage 2 estimates of V and 1.6 percentage points using Stage 1 estimates.

2. **Human information is valuable.** Accuracy under FDA significantly exceeds that achievable with AI alone ($p < 0.001$). The estimated difference is 1.6%, whereas the predictions are 1.5% and 1.8% using Stage 2 and Stage 1 estimates respectively.
3. **Human-AI collaboration does not outperform selective automation.** Accuracy under FDA does not significantly exceed that under NDA ($p = 0.44$). Human-AI collaboration increases accuracy by 0.2%, and our predictions using either estimate of V is within 0.2 percentage points of this estimate.
4. **AI assistance is valuable when automation is infeasible.** Accuracy under FDNA significantly exceeds that under NDNA ($p < 0.001$). We estimate an improvement of 3.4 percentage points from AI assistance without automation, as opposed to predictions of 5.4 percentage points and 6.2 percentage points from Stage 2 and Stage 1 respectively.
5. **Simple disclosure policies are approximately optimal.** Accuracy under FDNA does not significantly exceed that under SL ($p = 0.724$). Our experimental estimates suggest a small gain of 0.2 percentage points from using SL over FDA, whereas our predictions suggest a loss of 0.3 percentage points using either estimate of V .

These qualitative and quantitative conclusions are all robust to using an across-participant comparison based on the first treatment participants encounter, including controls for the treatment order or the number of prior statements assessed by the participant, or including participant fixed effects (see Table A.4).

As these qualitative conclusions were based on predictions about counterfactual accuracy made using Assumption 1, they represent a strong test of the sufficient-statistic approach.

There are, however, some departures from the more stringent standard of matching the model’s quantitative predictions. In particular, we correctly predict the quantitative value of automation (prediction 1) relative to the Stage 2 estimate of V but not relative to the Stage 1 estimate; and we mis-predict the quantitative value of AI assistance when automation is infeasible (prediction 4) for either the Stage 1 or Stage 2 estimate of V .

Table 2 provides results on the specific policies where our predictions do not match the experimental estimates. Columns 3 and 5 show that we cannot reject that accuracy under FDA, NDA, or SL equals the predicted accuracy using either estimate of V . However, accuracy under FDNA is 1.2 percentage points lower ($p = 0.008$) than accuracy from Stage

Table 2: Estimated Versus Predicted Accuracy

Treatment	Stage 2 Estimate	Stage 2		Stage 1	
		Predicted	P-value	Predicted	P-value
	(1)	(2)	(3)	(4)	(5)
<i>Panel A:</i>					
Full Disclosure + No Automation (FDNA)	0.723 (0.004)	-	-	0.735 (0.003)	0.014
<i>Panel B: Automation</i>					
Full Disclosure (FDA)	0.749 (0.002)	0.748 (0.003)	0.783	0.751 (0.002)	0.368
No Disclosure (NDA)	0.747 (0.001)	0.744 (0.003)	0.265	0.748 (0.002)	0.748
<i>Panel C: No Automation</i>					
No Disclosure (NDNA)	0.689 (0.004)	0.669 (0.009)	0.033	0.673 (0.006)	0.025
Stoplight (SL)	0.725 (0.004)	0.720 (0.006)	0.473	0.732 (0.004)	0.192
Joint Test	–	–	0.227	–	0.007

Note: In column (1), Stage 2 Estimate is the estimated accuracy from Stage 2 data. In column (2), Predicted is the predicted accuracy computed from Stage 2 data. In column (4), Predicted is the predicted accuracy computed from Stage 1 data, except for the FDNA row, which contains the observed accuracy in Stage 1. Columns (3) and (5) contain the P-value from the test where the null sets the Predicted and Stage 2 Estimate values to be the same. Standard errors are in parentheses. Predicted standard errors are computed via block bootstrap clustered at the participant level, and the Stage 2 Estimate standard errors are two-way clustered at the participant and case level. The Stage 2 p -values are based on a block bootstrap clustered at the participant level.

1; and accuracy under NDNA is 2.0 and 1.6 percentage points higher than the predicted accuracy using the estimate of V from Stage 2 and Stage 1, respectively (p -values < 0.05).

There are two distinct reasons why the predictions from the two stages may miss the experimental estimates. The first is that details of the experimental protocol might affect our participants' performance. For instance, there may be subtle differences in participants between the two stages, participants may learn how to use the AI differently in the two stages, and there may be effects of participants being exposed to multiple treatments in Stage 2. Figure 9a shows that the estimates of V from the two stages are similar—we cannot reject that the function V estimated in Stage 1 equals the same function estimated using Stage 2 FDNA data ($p = 0.285$). However, panel A of Table 2 shows that accuracy under FDNA is

lower in Stage 2 (72.3%) than in Stage 1 (73.5%).³² These differences do not imply a violation of Assumption 1, but they may explain why the predicted value of automation based on the Stage 1 estimate of V is quantitatively inaccurate, while the prediction based on the Stage 2 estimate is accurate.

The second reason is a violation of Assumption 1: participants’ accuracy may not depend only on the mean AI assessment. Participants’ greater accuracy under NDNA in Stage 2 relative to the model’s predictions may reflect such a violation. Specifically, a likely explanation for this finding is that cases where the AI is confident are also easier for human participants, so that participants’ average accuracy under NDNA is better than their accuracy on cases with the average AI assessment ϕ . Figure 9b points to this hypothesis. It plots participant accuracy as a function of the AI assessment θ under NDNA, where θ is *not* disclosed. It can be shown that Assumption 1 implies that this accuracy curve must be linear in θ for a Bayesian decision-maker, so Figure 9b suggests a likely violation of Assumption 1. However, the magnitude of the violation is small: participants’ average accuracy under NDNA is 1.6 percentage points higher than their accuracy on cases with the average AI assessment ϕ , suggesting that cases where the AI is more confident are only slightly easier for human participants. (To benchmark this number, note that predicted accuracy under the “opposite” assumption that human and AI signals are perfectly correlated is 73.5%—i.e., the same prediction as under FDNA—which exceeds actual accuracy under NDNA by a much larger 4.6 percentage points.)

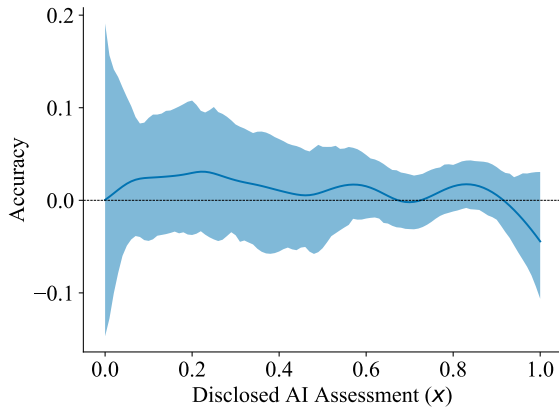
Another way to evaluate the sufficient-statistic approach is to compare the predicted accuracy at specific posteriors x to estimates from Stage 2. These estimates are displayed in Figure 9c, which shows that realized Stage 2 accuracy at the induced posteriors under NDA and NDNA—as well as at each of the three induced posteriors under SL—all closely match the corresponding values of $V(x)$ from the Stage 1 estimate of the function V . Our accuracy predictions are thus on point for each induced posterior, not just on average.

Overall, while predicted accuracy differs somewhat from estimated accuracy in a couple treatments, the sufficient-statistic approach based on Assumption 1 provides a useful guide to designing automation/disclosure policies in our setting.

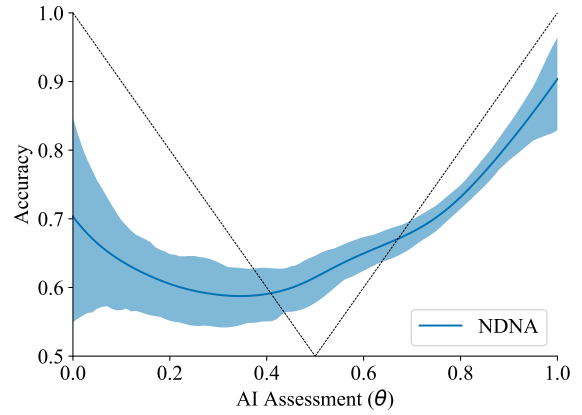
³²The difference is statistically significant ($p = 0.014$). Participants were also faster in Stage 2 than Stage 1. This can be explained by participants assessing more statements in Stage 2 while their assessments become faster over time (see Appendix A.3).

Figure 9: Stability

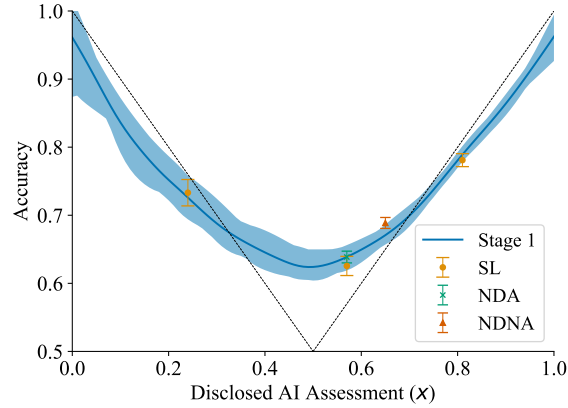
(a) Difference between Stage 1 and 2 Accuracy



(b) Accuracy as a Function of θ under NDNA

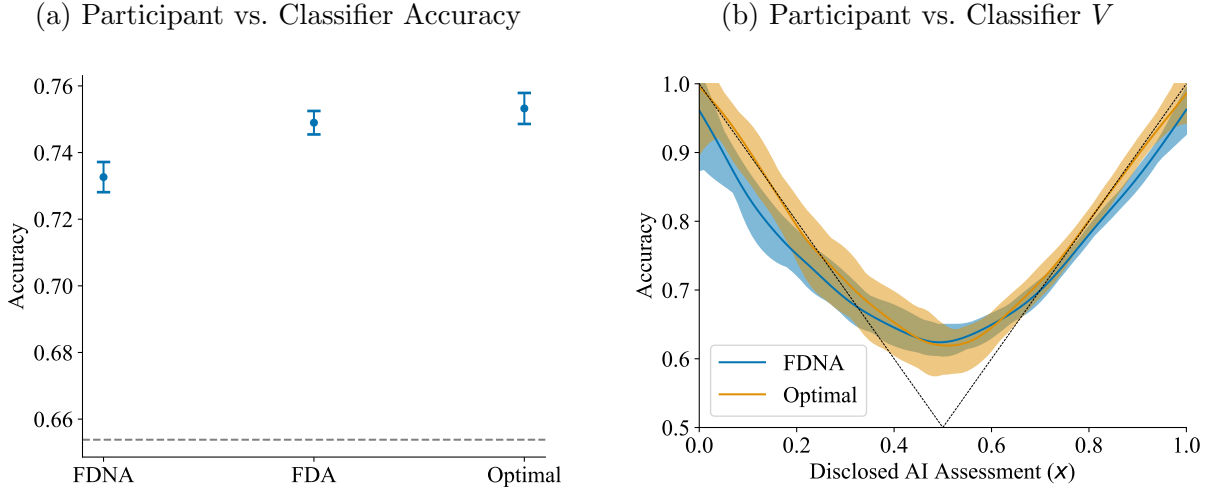


(c) Predicted vs Estimated Accuracy



Note: Figure 9a plots the difference in accuracy between Stage 1 and Stage 2, $V_1(x) - V_2(x)$, estimated via local linear regression separately for Stage 1 and Stage 2. Figure 9b plots accuracy conditional on θ for Stage 2 under the NDNA treatment estimated by local linear regression. Figure 9c plots accuracy conditional on x for Stage 1 via the function V estimated by local linear regression. The dashed lines indicate the accuracy under automation of $\max\{x, 1-x\}$. Accuracy by x under each policy is estimated by regressing a correct indicator on indicators for each AI assessment shown. The 95% point-wise confidence intervals for each point are two-way clustered at the participant and case level. For all three figures, the 95% uniform confidence band is computed via bootstrap accounting for clustering at the participant and case level.

Figure 10: Comparing Participant and Optimal Classifier Accuracy



Notes: Panel (a) plots accuracy under FDNA, FDA, and the optimal classifier. The horizontal dashed line is the accuracy of a classifier with no information that classifies all statements as True. Panel (b) plots the Stage 1 estimate of V and the estimated accuracy of the optimal classifier V^{Opt} .

5.2 Optimal Classifier Benchmark

We now calculate the accuracy of an optimal classifier V^{Opt} that uses both participants' reported assessments p_{ij} under FDNA and the AI assessments θ_i to classify each case i .³³ The optimal classification for case i is True if and only if $\Pr(\omega = 1 | p_{ij}, \theta_i)$ exceeds 0.5. We nonparametrically estimate $\Pr(\omega = 1 | p, \theta)$ as a function of (p, θ) using the FDNA sample from both stages. To avoid overfitting, we use a penalized logistic regression with polynomial terms in p and θ , where we use cross-validation to select the model to minimize expected out-of-sample loss.³⁴ Figure 10a compares accuracy under the tested policies to the optimal classifier V^{Opt} .

Our first result from this exercise is that average accuracy under the optimal classifier (75.3%) is statistically indistinguishable from accuracy under FDA (75.0%). This result implies that one-way communication from AI to humans is without loss of optimality: the optimal policy in our design space with no elicitation of participants' assessments (FDA) cannot be significantly improved by eliciting participants' assessments. Intuitively, this is a consequence of two properties of the classifier indirect utility function V^{Opt} , which is plotted alongside the human indirect utility function V in Figure 10b. First, V^{Opt} is indistinguishable from human accuracy V for AI assessments where delegation to a human is optimal (i.e., where

³³Guo et al. (2025) uses a related approach to measure the additional information contributed by an AI system over and above the information contained in humans' decisions.

³⁴Appendix E contains full details of the estimation.

$V(\theta) \geq \max\{\theta, 1 - \theta\}$). Second, $V^{\text{Opt}}(\theta)$ is indistinguishable from AI accuracy $\max\{\theta, 1 - \theta\}$ for AI assessments where automation is optimal (i.e., where $V(\theta) < \max\{\theta, 1 - \theta\}$). Together, these properties imply that selective automation achieves the optimal classifier benchmark.

Our second result is that accuracy under the optimal classifier is significantly greater than that under FDNA. This result implies that the impact of participants’ under-response to AI on accuracy is substantial. If participants were Bayesians who knew the joint distribution of (p, θ, ω) , their accuracy would be at least as high as the optimal classifier benchmark, because participants’ know their own assessments and the AI assessment and may also have additional information. This comparison gives a lower bound for the impact of non-Bayesian updating on participant accuracy. Thus, at least $(75.32\% - 73.26\%)/(1 - 73.26\%) = 7.7\%$ of incorrect classifications under FDNA are attributable to deviations from Bayesian updating.

Section 6 further unpacks the deviations from Bayesian updating that are responsible for this result.

5.3 Impact on Effort

Table 3 presents estimated treatment effects on our three measures of participant effort, relative to the baseline FDNA policy. It uses estimates from the model in equation (5) and reports β_{k_0} for FDNA and $\beta_k - \beta_{k_0}$ for the remaining policies.

Disclosing AI assessments crowds out human effort, consistent with Figure 7: our effort measures are between 8% and 13% lower under FDNA as compared to NDNA.³⁵ While this effort response is substantial, it is smaller than some related estimates in the literature: for example, Dell’Acqua (2022) finds that disclosing more precise AI assessments reduced effort by nearly 40%.

As with the estimated treatment effects on accuracy, this result is robust to a number of variations in the analysis (see Tables A.5, A.6, A.7, and A.8). When using the across-participant design based on the first treatment encountered—Table A.5—the treatment effects are similar to the within-participant design, except the baseline for measures of effort is higher. This can be explained by the fatigue effects presented in Table A.3.

The next section presents additional results on the impact of effort crowd-out on the informativeness of humans’ signals.

³⁵Effort under Stoplight is indistinguishable from that under FDNA, consistent with these two policies being very similar

Table 3: Average Treatment Effects on Effort

Treatment	External Sources	Clicked Google	Time Taken (s)
	(1)	(2)	(3)
<i>Panel A: No Automation Baseline (β_0)</i>			
Full Disclosure	0.630 (0.009)	0.372 (0.009)	44.551 (0.730)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
Full Disclosure	-0.357 (0.007)	-0.209 (0.007)	-24.515 (0.560)
No Disclosure	-0.412 (0.007)	-0.240 (0.007)	-28.523 (0.586)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
No Disclosure	0.064 (0.006)	0.046 (0.006)	3.749 (0.586)
Stoplight	0.003 (0.005)	0.002 (0.006)	0.091 (0.529)
Observations	80000	80000	80000

Note: Average treatment effects (estimated using equation 5) of different polices on effort. In FDA and NDA, outcomes have been adjusted to account for automation as described in Footnote 31.

6 Mechanisms: Overconfidence, AI Neglect, and Effort Crowd-Out

The estimates presented in Sections 4 and 5 show that our participants under-respond to AI assessments and reduce effort when presented with confident AI assessments. This section analyzes participants’ biases in belief updating and the impact of effort crowd-out on accuracy.³⁶

Specifically, we distinguish between participants’ *overconfidence* in the precision of their own information and *under-confidence* in the precision of AI information—which we refer to as *AI neglect*. Empirically distinguishing overconfidence from AI neglect requires additional assumptions to nonparametrically identify the distribution of participants’ private information and their model of belief updating. Under these assumptions, we also show that the

³⁶The analyses in this section were not pre-registered.

reduction in participant accuracy due to measured effort crowd-out is modest in magnitude.

6.1 Over- and Under-Inference

We will use the following general definition of over- or under-inference from a signal to define overconfidence and AI neglect. Consider an agent who observes a vector of N real-valued signals $\mathbf{s} = (s_1, \dots, s_N) \in \mathbb{R}^N$ of a binary state $\omega \in \{0, 1\}$. Assume that each signal s_n is ordered by likelihood ratios, so that $\Pr(s_n = s | \omega = 1) / \Pr(s_n = s | \omega = 0)$ is increasing in $s \in \mathbb{R}$. For example, this property holds if each signal s_n is calibrated (i.e., $s_n \in [0, 1]$ and $\Pr(\omega = 1 | s_n = s) = s$ for all $s \in [0, 1]$).³⁷

Let $p(\mathbf{s}) \in [0, 1]$ denote the agent's assessment of the probability that $\omega = 1$ at signal vector \mathbf{s} , and let $p^{\text{Bayes}}(\mathbf{s}) = \Pr(\omega = 1 | \mathbf{s})$ be the corresponding Bayesian assessment. We say that the agent *over-infers* from a signal s_n if the proportional increase in her posterior odds ratio of $\omega = 1$ to $\omega = 0$ from observing a higher signal s_n is always greater than that for a Bayesian: that is, if

$$\frac{p(s'_n, \mathbf{s}_{-n})}{1 - p(s'_n, \mathbf{s}_{-n})} \frac{1 - p(s_n, \mathbf{s}_{-n})}{p(s_n, \mathbf{s}_{-n})} > \frac{p^{\text{Bayes}}(s'_n, \mathbf{s}_{-n})}{1 - p^{\text{Bayes}}(s'_n, \mathbf{s}_{-n})} \frac{1 - p^{\text{Bayes}}(s_n, \mathbf{s}_{-n})}{p^{\text{Bayes}}(s_n, \mathbf{s}_{-n})},$$

for all $s'_n > s_n$ and all $\mathbf{s}_{-n} \in \mathbb{R}^{N-1}$ such that $0 < p(s_n, \mathbf{s}_{-n}) \leq p(s'_n, \mathbf{s}_{-n}) < 1$.

Similarly, the agent *under-infers* from s_n if the same condition holds with the reverse inequality. Note that if p is continuously differentiable then, letting $\text{logit } x = \log \frac{x}{1-x}$, an equivalent definition of over-inference from s_n is

$$\frac{\partial}{\partial s_n} \text{logit } p(s_n, \mathbf{s}_{-n}) > \frac{\partial}{\partial s_n} \text{logit } p^{\text{Bayes}}(s_n, \mathbf{s}_{-n}) \text{ for all } \mathbf{s} \in \mathbb{R}^N \text{ such that } 0 < p(\mathbf{s}) < 1. \quad (6)$$

As far as we know, this definition of over-inference is novel, although it has some close predecessors. The closest is in Augenblick et al. (2025), which defines a notion of the perceived strength $\hat{S}(s)$ of a signal s and say that an agent over-infers from s if they over-perceive the strength of s and then update according to Bayes' rule. With a single signal, Augenblick et al.'s definition appears to have the same implications for belief updating as ours, but we allow multiple signals and do not invoke the notion of perceived signal strength. Our definition also generalizes those in Grether (1980) and Agarwal et al. (2023). In particular, if signals are conditionally independent and calibrated, and we rewrite (6) in an equivalent form where the derivatives are taken with respect to $\text{logit } s_n$ rather than s_n , then the right-hand side is

³⁷In this case, $\frac{\Pr(s_n = s | \omega = 1)}{\Pr(s_n = s | \omega = 0)} = \frac{1 - \phi}{\phi} \frac{\Pr(\omega = 1 | s_n = s)}{\Pr(\omega = 0 | s_n = s)} = \frac{1 - \phi}{\phi} \frac{s}{1 - s}$, where $\phi = \Pr(\omega = 1)$.

1 and the left-hand side is the Grether coefficient on signal s_n .³⁸

In our setting, humans obtain two calibrated signals—a private signal s (described below) and the disclosed AI assessment x —and combine them to form an assessment $f(s, x)$. Applying the above general definition, we say that humans are *overconfident* in their own signal if they over-infer from s , and that they display *AI neglect* if they under-infer from x .

6.2 Identifying Participant Signals and Updating

The following assumptions let us identify participants’ signals and belief updating model.

Assumption 2.1 *Humans observe a one-dimensional signal $s_{ij} \in [0, 1]$ that is distributed iid conditional on $\omega_i, e_{ij}, \theta_i$ with cumulative distribution function (CDF) $G_{\omega_i, e_{ij}, \theta_i}$, where e_{ij} is the vector of observed measures of effort. Without loss of generality, we normalize $s_{ij} = P(\omega_i = 1 | s_{ij})$, so the human signal is calibrated.*

Assumption 2.2 *Humans’ reported assessments p_{ij} are determined by their own signals s_{ij} and the disclosed AI assessments x_i according to a function $p(s_{ij}, x_i) = p_{ij}$, which is monotone in s_{ij} .*

Assumption 2.1 imposes two restrictions. First, the distribution of human signals does not depend on the disclosure policy or the disclosed AI signal x_i conditional on $\omega_i, e_{ij}, \theta_i$.³⁹ In particular, our observed measures of effort e_{ij} —time taken, an indicator for the reported use of external sources, and an indicator for clicking the Google search link—are sufficient controls for the dependence of the human signal s_{ij} on the disclosed AI signal x_i . Second, while the distribution of effort can vary across human participants, the signal distribution is the same across participants conditional on effort.

Assumption 2.2 imposes three restrictions. First, the human assessment p_{ij} depends only on the human signal s_{ij} and the disclosed AI assessment x_i and not on other observables (such as effort e_{ij}). Second, the assessment is monotone in the human signal.⁴⁰ For example, Assumption 2.2 holds if humans are Bayesian with conditionally independent signals. It also holds if humans are quasi-Bayesians who act as if their signals are conditionally independent

³⁸In the conditionally independent case, the models in Grether (1980) and Agarwal et al. (2023) assume that $\text{logit } p(\mathbf{s}) = \sum_{n=1}^N a_n (\text{logit } \Pr(\omega = 1 | s_n) - \text{logit } \Pr(\omega = 1)) + b \text{logit } \Pr(\omega = 1)$ for parameters a_1, \dots, a_N, b . For $p^{\text{Bayes}}(\cdot)$, we have that $a_1 = \dots = a_N = b = 1$.

³⁹We allow for dependence on θ_i because the AI assessment can be statistically dependent. The distribution of signals can also depend on the disclosure policy or the disclosed signal, but only via observed effort.

⁴⁰It is natural to assume that the assessment is also monotone in the disclosed AI assessment x_i , but our identification strategy does not require this assumption.

of the AI signal and may over- or under-weight either signal, as in Grether (1980) or (Agarwal et al., 2023). Third, the function $p(\cdot)$ is the same for all participants.

Assumptions 2.1 and 2.2 allow us to identify and estimate $p(\cdot)$. We first explain how to calculate $p(s, x)$ at s and $x = \theta$ from the conditional CDFs of human assessments p and human signals s given each AI assessment θ under FDNA, which we denote by $F_{p|\theta}$ and $F_{s|\theta}$, and then explain how we identify and estimate these CDFs. By Assumption 2.2, for any human signal s and AI signal θ in FDNA, we have

$$F_{s|\theta}(s) = F_{p|\theta}(p(s, \theta)).$$

Thus, inverting the CDF $F_{p|\theta}$ gives

$$p(s, \theta) = F_{p|\theta}^{-1}(F_{s|\theta}(s)). \quad (7)$$

The conditional CDF $F_{p|\theta}$ is observed under FDNA and we estimate it nonparametrically.⁴¹ The remaining task is to identify and estimate $F_{s|\theta}$. We accomplish this in two steps.

First, we identify and estimate the human signal distribution $G_{\omega_i, e_{ij}, \theta_i}$ using data from NDNA. By Assumption 2.1, $G_{\omega_i, e_{ij}, \theta_i}$ is independent of the disclosure policy and the disclosed AI assessment x_i , conditional on $(\omega_i, e_{ij}, \theta_i)$. Under NDNA, the disclosed AI assessment x_i is constant at the prior ϕ , while participants report continuous assessments p_{ij} of the probability that each statement i is true. Since x_i is constant, Assumption 2.2 implies that p_{ij} is a deterministic function of s_{ij} , and hence $\Pr(\omega_i = 1 | p_{ij}) = \Pr(\omega_i = 1 | s_{ij}) = s_{ij}$. Thus, under NDNA s_{ij} , ω_i , e_{ij} , and θ_i are observable, and hence we can identify and estimate $G_{\omega_i, e_{ij}, \theta_i}$ nonparametrically (see footnote 41).

Next, the conditional CDF $F_{s|\theta}$ can be calculated from $G_{\omega_i, e_{ij}, \theta_i}$ identified from the NDNA data by integrating over the observed joint distribution of ω_i and e_{ij} in FDNA. We estimate this distribution fitting a conditional distribution model (see footnote 41) to 100,000 simulated draws from the joint distribution of s_{ij} , e_{ij} , θ_i , and ω_i in the FDNA arm. To generate these draws, we first sample from the joint distribution of e_{ij} , θ_i , and ω_i using an accept/reject

⁴¹We estimate all conditional CDFs of the form $F_{y|\theta}(z)$ using a logistic regression of the indicator $1[y \leq z]$ on θ including second-order polynomials and all second-order interactions when z is a vector. We estimate this for a grid of z to trace out the full conditional CDF (Chernozhukov et al., 2013). When the CDF is non-monotonic we apply the rearrangement procedure described in Chernozhukov et al. (2010).

sampler.^{42,43} Then, we sample s_{ij} from the conditional distribution $G_{\omega_i, e_{ij}, \theta_i}$ estimated earlier using inverse transform sampling.

Finally, we use a plug-in estimator that replaces the conditional distributions of p and s with the estimated analogues in equation (7).⁴⁴

6.3 Overconfidence or AI Neglect?

Having estimated our participants’ belief updating rule $p(s, x)$, we can now compare it to the Bayesian benchmark $p^{\text{Bayes}}(s, x)$ to decompose the AI under-response found in Section 4 into overconfidence in participants’ own signals and AI neglect.

We estimate $p^{\text{Bayes}}(s, x)$ through a penalized logistic regression of ω on s and x in the 100,000 samples of ω , s , and θ from the FDNA arm described above (see Appendix E for estimation details).

Figure 11 presents estimates of our participants’ update function p (blue curve) and the Bayesian benchmark rule p^{Bayes} (orange curve), as well as the Bayesian benchmark imposing conditional independence (green curve; a line of slope 1 in log odds space). The panels hold either s or x fixed at a specific value while varying the other signal in log odds space. (Appendix Figure C.7 presents a corresponding surface plot.) A first observation is that the two Bayesian benchmarks are quite similar, implying that conditional independence is a reasonable approximation.

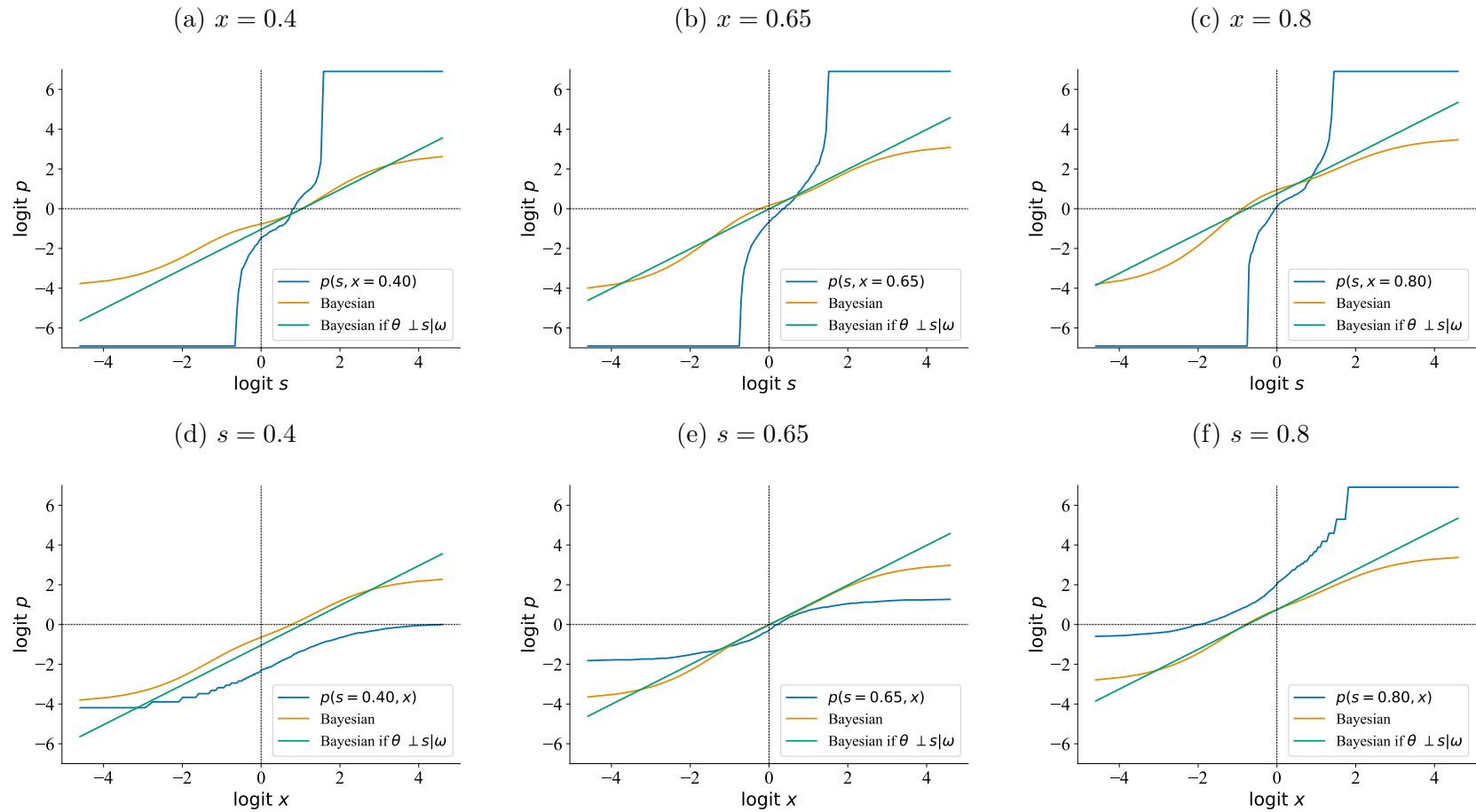
Figures 11a–11c show strong evidence of overconfidence. Recall that, as defined in Section 6.1, overconfidence means that $\text{logit } p$ is steeper than $\text{logit } p^{\text{Bayes}}$ when s varies, and AI neglect means that $\text{logit } p$ is flatter than $\text{logit } p^{\text{Bayes}}$ when θ varies. Correspondingly, the slope of $\text{logit } p$ with respect to s in Figures 11a–11c is much larger than the Bayesian benchmark. This overconfidence results in participants reporting more extreme probability assessments than a calibrated decision maker (as shown in Figure 6), as well as in AI under-response (e.g., $V(x) < \max\{x, 1 - x\}$ for x near 0 or 1).

⁴²We estimate the joint distribution of e_{ij} , θ_i , and ω_i in the FDNA arm using kernel density estimation. We use a Gaussian kernel for all continuous variables and Silverman’s rule to select bandwidths (Silverman, 2018). We manually select a bandwidth of 0 for all binary variables.

⁴³In the FDNA arm, $P(\omega_i = 1) = 0.657$, while $P(\omega_i = 1) = 0.649$ in the NDNA arm. While we cannot reject that this difference is zero ($p = 0.14$), we sample from the population distribution of ω_i to impose balance.

⁴⁴This approach to identifying participants’ update rule has several advantages over the one in Agarwal et al. (2023). Agarwal et al. (2023) requires participants to assess the same case twice, once with AI assistance and once without. In addition, our approach allows (observed) effort responses to influence the signal distribution. However, we require human signals to be one-dimensional.

Figure 11: Human vs Bayesian Update Rule



Note: This figure summarizes the human and Bayesian update rules. Panels (d)-(f) plot $p(s, x)$ and $P(\omega = 1 | x, s)$ for different values of s and Panels (a)-(c) plot these functions for different values x . All figures are in log-odds space.

In contrast, Figures 11d–11f show weaker evidence of AI neglect: the slopes of logit f and logit f^{Bayes} with respect to θ are fairly similar, although logit f is somewhat flatter, indicating some degree of AI neglect. The vertical shifts in the logit $f(s, \cdot)$ curve relative to logit $p^{\text{Bayes}}(s, \cdot)$ at $s = 0.4$ and $s = 0.8$ reflect overconfidence.

Overall, Figure 11 shows strong evidence of overconfidence, as well as some evidence of AI neglect.

Next, we quantify the relative impact of overconfidence and AI neglect by comparing the accuracy of a decision-maker who exhibits only automation neglect or only overconfidence. To do so, we define the human assessment corrected for overconfidence as $\tilde{p}(s, x)$ such that

$$\text{logit } \tilde{p}(s, x) = \text{logit } p^{\text{Bayes}}(s, x) + \text{logit } p(\phi, x) - \text{logit } p^{\text{Bayes}}(\phi, x),$$

where $\phi = \Pr(\omega = 1)$ is the prior mean. Here, $\frac{\partial}{\partial s} \text{logit } \tilde{p} = \frac{\partial}{\partial s} \text{logit } p^{\text{Bayes}}$, so \tilde{p} and the Bayesian benchmark respond equally to changes in the human signal, which removes overconfidence. The remaining terms are set so that $\tilde{p}(\phi, \phi) = p(\phi, \phi)$ to ensure that \tilde{p} matches the human assessment when s and x are uninformative; and $\frac{\partial}{\partial x} \text{logit } \tilde{p}(\phi, x) = \frac{\partial}{\partial x} \text{logit } p(\phi, x)$ to ensure that \tilde{p} and the human assessment respond equally to changes in x when s is uninformative. Similarly, we define the human assessment corrected for AI neglect as $\check{p}(s, x)$ such that

$$\text{logit } \check{p}(s, x) = \text{logit } p^{\text{Bayes}}(s, x) + \text{logit } p(s, \phi) - \text{logit } p^{\text{Bayes}}(s, \phi).$$

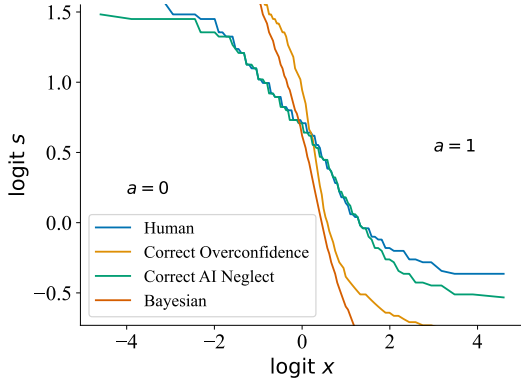
Figure 12a plots the decision threshold in $(\text{logit } x, \text{logit } s)$ -space for humans, Bayesians, and humans corrected for overconfidence or AI neglect. We see that the decision threshold for overconfidence-corrected humans is very close to the Bayesian benchmark, while the threshold for AI neglect-corrected humans is very close to that for uncorrected humans. Correspondingly, Figure 12b shows that correcting AI neglect increases accuracy by only 0.1 percentage points, while correcting overconfidence increases accuracy by 1.7 percentage points (out of a possible improvement of 2.2 percentage points for the Bayesian benchmark). These results show that overconfidence—not AI neglect—is the main reason our participants deviate from optimal Bayesian decisions.

Our result that AI under-response is primarily due to overconfidence rather than AI neglect differs from that in Agarwal et al. (2023), which finds evidence for AI neglect but not overconfidence among professional radiologists.⁴⁵ One possible hypothesis for this difference is that professional decision-makers (e.g., the radiologists in Agarwal et al. (2023)) understand

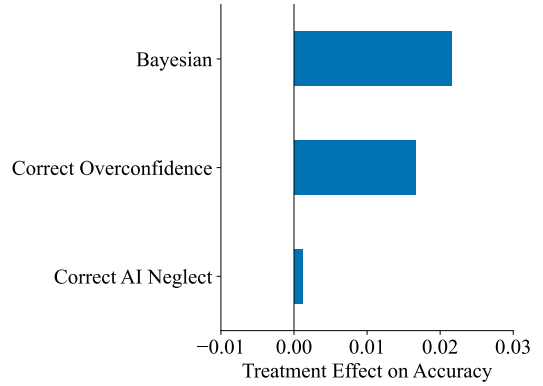
⁴⁵Agarwal et al. (2023) estimates the Grether model where $\text{logit } p(s, x) = a + b \text{logit } s + c \text{logit } x$, finding that $b = 0.3$ and $c = 1.1$. In contrast, estimating the same model with our data yields $b = 0.8$ and $c = 2.3$.

Figure 12: Decomposing Overconfidence and AI Neglect

(a) Impact of Biases on Decision Threshold



(b) Impact of Biases on Accuracy



Note: Panel (a) plots the decision threshold for various decision-makers. Each curve is the set of points (s, x) where $p(s, x) = 0.5$ for each decision-maker. The range of the y-axis is the support of logit s . Panel (b) plots the accuracy of each decision-maker relative to human participants.

their own abilities but distrust outside advice, while amateurs (e.g., our participants) overestimate their own abilities but are more open to advice.

6.4 Impact of Effort Crowd-Out on Human Signal Quality

Our identification of $G_{\omega_i, e_{ij}, \theta_i}$ under Assumptions 2.1 and 2.2 also lets us measure the impact of effort crowd-out on the precision of human signals. Specifically, we use our estimate of $G_{\omega_i, e_{ij}, \theta_i}$ to compare the quality of the human signal s under FDNA and NDNA for various ranges of the AI assessment θ : $\theta < 0.25$, $\theta \in [0.25, 0.75]$, and $\theta > 0.75$. Table 4 presents the treatment effect of disclosure on our observed measures of effort and human signal precision calculated using our estimate of $G_{\omega_i, e_{ij}, \theta_i}$.

Panel A shows that disclosing the AI assessment reduces our three effort measures for all AI assessment ranges. The decline in effort is much larger when the AI is confident ($\theta < 0.25$ or $\theta > 0.75$). This is intuitive and is consistent with the overall treatment effects on effort documented in Section 5.3.

Panel B shows that this effort crowding-out also reduces three measures of human signal precision. The first row shows that effort crowding-out increases the root mean-square error of the human signal. The second row shows that it reduces the probability that the human signal alone would result in a correct classification. The third row shows that it increases the probability that the human signal is insufficient to overturn the prior favoring classifying cases as True.⁴⁶ All of these reductions in precision are concentrated on statements that the

⁴⁶Recall that 65.4% of cases in the database are true.

Table 4: Impact of Disclosing AI Assessment on Effort Human Signal Precision

	$\theta < 0.25$	$\theta \in [0.25, 0.75]$	$\theta > 0.75$	All Statements
<i>Panel A: Effort Measures</i>				
External Sources	-0.074 (0.019)	-0.029 (0.008)	-0.106 (0.009)	-0.064 (0.006)
Clicked Google	-0.039 (0.019)	-0.019 (0.008)	-0.080 (0.010)	-0.046 (0.006)
Page Time (Seconds)	-4.361 (1.686)	-1.033 (0.759)	-7.037 (0.823)	-3.749 (0.586)
<i>Panel B: Human Signal</i>				
RMSE	0.010 (0.004)	-0.001 (0.001)	0.004 (0.002)	0.001 (0.001)
Pr(Correct s_{ij})	-0.027 (0.013)	0.002 (0.003)	-0.003 (0.004)	-0.001 (0.002)
Pr(True s_{ij})	0.023 (0.014)	0.001 (0.003)	0.007 (0.003)	0.004 (0.002)

Note: Impact of FDNA relative to NDNA on effort and the precision of the human signal s . We report all measures averaging over all statements as well as conditional on the AI assessment θ . Panel (a) reports differences in participant effort under FDNA relative to NDNA. Panel (b) reports the treatment effect of FDNA relative to NDNA on the root mean squared error of the human signal (RMSE = $\left(\mathbb{E} \left[(\Pr(\omega = 1|s) - \omega)^2 \right] \right)^{1/2}$), the probability of correctly classifying a statement based on the human signal ($\Pr(\text{Correct}) = \Pr(1[\Pr(\omega = 1|s) > 1/2] = \omega)$), and the probability of classifying a statement as True based on the human signal ($\Pr(\text{True}) = \Pr(\Pr(\omega = 1|s) > 1/2)$). Bootstrapped standard errors in parenthesis.

AI is confident are false ($\theta < 0.25$). A possible explanation for the asymmetry between cases where $\theta < 0.25$ and where $\theta > 0.75$ is that, since cases where $\theta < 0.25$ are rare (see Figure 4), disclosing that $\theta < 0.25$ has a larger effect on participant effort and beliefs.

Overall, Table 4 provides modest evidence that effort crowding-out due to AI disclosure reduces human signal precision and contributes to the value of selective automation. However, the effect sizes are much smaller than what might be expected from other studies (e.g., Dell’Acqua (2022)).

7 Conclusion

Collaboration between humans and AI already profoundly affects organizational decision-making and job design, and its importance will only grow over time (Daugherty and Wilson, 2018; Mollick, 2024). The design of effective human-AI collaborative systems is thus a pressing concern. The standard approach to this problem in the current literature is “algorithmic

triage” (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023), which focuses on deciding which cases to automate and which to assign to humans, with or without AI assistance. However, this approach does not allow richer designs that partially disclose AI information, and it also does not account for the endogenous response of human beliefs and effort to the set of cases that are assigned to humans and the AI assistance provided. Moreover, the dimensionality of the space of possible collaborative designs and the complexity of possible human responses imply that optimal designs cannot be found by straightforward experimentation.

Our contribution is to show that, for binary classification problems, the optimal design can be found by estimating a simple sufficient statistic: the probability of correct classification as a function of the disclosed posterior. We validate this approach in the context of an online fact-checking experiment, where we show that the optimal policy automates cases where the AI is confident and delegates the remaining cases to human decision-makers while fully disclosing the AI assessment. At the same time, even simpler policies—such as selective automation without direct human-AI communication—are approximately optimal. We also show that the value of automation stems from human under-response to AI information, which in turn results from human over-confidence in the precision of their own information, rather than under-confidence in the AI.

A promising avenue for future research is to broaden the scope of human-AI collaboration design beyond binary classification and prediction problems—while this class of problems includes many important examples, extending the approach to higher-dimensional prediction problems would also be valuable. As discussed in Section 2, this extension would build on the general information design framework in Kamenica and Gentzkow (2011), rather than the more specialized framework in Dworzak and Martini (2019) that we utilize. More broadly, designing human-AI collaboration for problems other than prediction is another open area.

The space of collaborative policies considered can also be enlarged. For example, while we document substantial effort response to AI information disclosure, we do not consider the joint design of an information disclosure policy and an incentive contract. Similarly, we document significant biases in belief updating in response to AI information, but we do not consider policies targeting at reducing these biases, such as training humans to put more weight on AI information or less weight on their own information.

In addition to designing human-AI collaboration, our sufficient statistic can also be used to evaluate changes in the quality of AI information. In our framework, changing the underlying predictive AI tool corresponds to changing the distribution F over AI assessments θ . It is thus straightforward to calculate how changes in the AI affect the optimal collaborative policy

and the resulting decision accuracy. We leave this direction for future work.

Finally, we consider a setting where the statements to be classified are politically neutral, and the designer's objective of maximizing classification accuracy is aligned with the agent's (except for effort costs borne by the agent). An interesting avenue of research is designing AI information provision to persuade agents who may have misaligned objectives or motivated beliefs. This case may be relevant for fact-checking politically charged statements.

References

- Agarwal, Nikhil, Alex Moehring, Pranav Rajpurkar, and Tobias Salz**, “Combining human expertise with artificial intelligence: Experimental evidence from radiology,” Technical Report, National Bureau of Economic Research 2023.
- Agrawal, Ajay, Joshua Gans, and Avi Goldfarb**, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press, April 2018.
- Allen, Jennifer, Antonio A Arechar, Gordon Pennycook, and David G Rand**, “Scaling up fact-checking using the wisdom of crowds,” *Science advances*, 2021, 7 (36), eabf4393.
- Aly, Rami, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos Christodoulopoulos, Oana Cocarascu, and Arpit Mittal**, “Feverous: Fact extraction and verification over unstructured and structured information,” *arXiv preprint arXiv:2106.05707*, 2021.
- Angelova, Victoria, Will S Dobbie, and Crystal Yang**, “Algorithmic recommendations and human discretion,” Technical Report, National Bureau of Economic Research 2023.
- Arieli, Itai, Yakov Babichenko, Rann Smorodinsky, and Takuro Yamashita**, “Optimal persuasion via bi-pooling,” *Theoretical Economics*, 2023, 18 (1), 15–36.
- Athey, Susan C, Kevin A Bryan, and Joshua S Gans**, “The allocation of decision authority to human and artificial intelligence,” in “AEA Papers and Proceedings,” Vol. 110 American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203 2020, pp. 80–84.
- Augenblick, Ned, Eben Lazarus, and Michael Thaler**, “Overinference from weak signals and underinference from strong signals,” *The Quarterly Journal of Economics*, 2025, 140 (1), 335–401.
- Benjamin, Daniel**, “Chapter 2 - Errors in probabilistic reasoning and judgment biases,” *Handbook of Behavioral Economics: Applications and Foundations 1*, January 2019, 2, 69–186. <https://doi.org/10.1016/bs.hesbe.2018.11.002>.
- Blackwell, David**, “Equivalent Comparisons of Experiments,” *The Annals of Mathematical Statistics*, 1953, 24 (2), 265–272. <https://www.jstor.org/stable/2236332>.

- Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond**, “Generative AI at work,” *The Quarterly Journal of Economics*, 2025, p. qjae044.
- Chan, David C, Matthew Gentzkow, and Chuan Yu**, “Selection with variation in diagnostic skill: Evidence from radiologists,” *The Quarterly Journal of Economics*, 2022, 137 (2), 729–783.
- Chen, Daniel L, Martin Schonger, and Chris Wickens**, “oTree—An open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance*, 2016, 9, 88–97.
- Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon**, “Quantile and probability curves without crossing,” *Econometrica*, 2010, 78 (3), 1093–1125.
- , —, and **Blaise Melly**, “Inference on counterfactual distributions,” *Econometrica*, 2013, 81 (6), 2205–2268.
- Chetty, Raj**, “Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods,” *Annu. Rev. Econ.*, 2009, 1 (1), 451–488.
- Clippel, Geoffroy De and Xu Zhang**, “Non-bayesian persuasion,” *Journal of Political Economy*, 2022, 130 (10), 2594–2642.
- Daugherty, Paul R and H James Wilson**, *Human+ machine: Reimagining work in the age of AI*, Harvard Business Press, 2018.
- Dell’Acqua, Fabrizio**, “Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters,” Technical Report, Working Paper 2022.
- Dietvorst, Berkeley J, Joseph P Simmons, and Cade Massey**, “Algorithm aversion: people erroneously avoid algorithms after seeing them err.,” *Journal of experimental psychology: General*, 2015, 144 (1), 114.
- Dubé, Jean-Pierre and Sanjog Misra**, “Personalized pricing and consumer welfare,” *Journal of Political Economy*, 2023, 131 (1), 131–189.
- Dworzak, Piotr and Giorgio Martini**, “The simple economics of optimal persuasion,” *Journal of Political Economy*, 2019, 127 (5), 1993–2048. <https://doi.org/10.1086/701813>.
- Facebook**, “Third-Party Fact-Checking Program by Facebook,” <https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking> Accessed August 12, 2024.

- Fréchet**, **Guillaume R**, **Alessandro Lizzeri**, and **Jacopo Perego**, “Rules and commitment in communication: An experimental analysis,” *Econometrica*, 2022, *90* (5), 2283–2318.
- Gentzkow**, **Matthew** and **Emir Kamenica**, “A Rothschild-Stiglitz approach to Bayesian persuasion,” *American Economic Review*, 2016, *106* (5), 597–601.
- Grether**, **David M**, “Bayes rule as a descriptive model: The representativeness heuristic,” *The Quarterly journal of economics*, 1980, *95* (3), 537–557.
- Guo**, **Zhijiang**, **Michael Schlichtkrull**, and **Andreas Vlachos**, “A survey on automated fact-checking,” *Transactions of the Association for Computational Linguistics*, 2022, *10*, 178–206.
- Guo**, **Ziyang**, **Yifan Wu**, **Jason Hartline**, and **Jessica Hullman**, “The Value of Information in Human-AI Decision-making,” *arXiv preprint arXiv:2502.06152*, 2025.
- Hastie**, **Trevor**, **Robert Tibshirani**, and **Jerome Friedman**, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Springer Series in Statistics, 2 ed., New York: Springer, 2009.
- Hirano**, **Keisuke**, **Guido W Imbens**, and **Geert Ridder**, “Efficient estimation of average treatment effects using the estimated propensity score,” *Econometrica*, 2003, *71* (4), 1161–1189.
- Hossain**, **Tanjim** and **Ryo Okui**, “The binarized scoring rule,” *Review of Economic Studies*, 2013, *80* (3), 984–1001.
- International Fact-Checking Network**, “State of Fact-Checkers 2023,” Technical Report, Poynter Institute 2023.
- Kamenica**, **Emir** and **Matthew Gentzkow**, “Bayesian Persuasion,” *American Economic Review*, October 2011, *101* (6), 2590–2615. <https://www.aeaweb.org/articles?id=10.1257/aer.101.6.2590>.
- Kaplan**, **Joel**, “More Speech and Fewer Mistakes,” 2025. Accessed: 2025-02-26.
- Kleinberg**, **Jon**, **Himabindu Lakkaraju**, **Jure Leskovec**, **Jens Ludwig**, and **Sendhil Mullainathan**, “Human Decisions and Machine Predictions,” *Q. J. Econ.*, August 2017, *133* (1), 237–293.

- Kolotilin, Anton**, “Optimal information disclosure: A linear programming approach,” *Theoretical Economics*, 2018, *13* (2), 607–635. <https://doi.org/10.3982/TE1805>.
- , **Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li**, “Persuasion of a privately informed receiver,” *Econometrica*, 2017, *85* (6), 1949–1964. <https://doi.org/10.3982/ECTA13251>.
- Lai, Vivian, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan**, “Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies,” December 2021.
- Lazer, David MJ, Matthew A Baum, Yochai Benkler, Adam J Berinsky, Kelly M Greenhill, Filippo Menczer, Miriam J Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild et al.**, “The science of fake news,” *Science*, 2018, *359* (6380), 1094–1096.
- Li, Danielle, Lindsey R Raymond, and Peter Bergman**, “Hiring as exploration,” Technical Report, National Bureau of Economic Research 2020.
- Longoni, Chiara, Andrea Bonezzi, and Carey K Morewedge**, “Resistance to medical artificial intelligence,” *Journal of consumer research*, 2019, *46* (4), 629–650.
- Lowenkamp, Christopher T**, “The development of an actuarial risk assessment instrument for US Pretrial Services,” *Fed. probation*, 2009, *73*, 33.
- Maslej, Nestor, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark**, “The AI Index 2024 Annual Report,” Technical Report, AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA April 2024.
- Misra, Sanjog and Harikesh S Nair**, “A structural model of sales-force compensation dynamics: Estimation and field implementation,” *Quantitative Marketing and Economics*, 2011, *9*, 211–257.
- Mollick, Ethan**, *Co-intelligence: Living and working with AI*, Penguin, 2024.
- Moore, Don A and Paul J Healy**, “The trouble with overconfidence.,” *Psychological review*, 2008, *115* (2), 502.

- Mozannar, Hussein and David Sontag**, “Consistent estimators for learning to defer to an expert,” in “International conference on machine learning” PMLR 2020, pp. 7076–7087.
- Mullainathan, Sendhil and Ziad Obermeyer**, “Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care,” *The Quarterly Journal of Economics*, May 2022, *137* (2), 679–727. <https://doi.org/10.1093/qje/qjab046>.
- Noy, Shakked and Whitney Zhang**, “Experimental evidence on the productivity effects of generative artificial intelligence,” *Science*, 2023, *381* (6654), 187–192. <https://doi.org/10.1126/science.adh2586>.
- Olea, José Luis Montiel and Mikkel Plagborg-Møller**, “Simultaneous confidence bands: Theory, implementation, and an application to SVARs,” *Journal of Applied Econometrics*, 2019, *34* (1), 1–17.
- Ostrovsky, Michael and Michael Schwarz**, “Reserve prices in internet advertising auctions: A field experiment,” *Journal of Political Economy*, 2023, *131* (12), 3352–3376.
- Policy, Advancing Pretrial and Research**, “Guide to the Release Conditions Matrix,” June 2020. Supported by Arnold Ventures.
- Raghu, Maithra, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan**, “The algorithmic automation problem: Prediction, triage, and human effort,” *arXiv preprint arXiv:1903.12220*, 2019.
- Silverman, Bernard W**, *Density estimation for statistics and data analysis*, Routledge, 2018.
- Skitka, Linda J, Kathleen L Mosier, and Mark Burdick**, “Does automation bias decision-making?,” *International Journal of Human-Computer Studies*, 1999, *51* (5), 991–1006.
- Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone**, “When Are Combinations of Humans and AI Useful?,” *arXiv e-prints*, 2024, pp. arXiv–2405.
- Vodrahalli, Kailas, Tobias Gerstenberg, and James Y Zou**, “Uncalibrated models can improve human-ai collaboration,” *Advances in Neural Information Processing Systems*, 2022, *35*, 4004–4016.
- X Community Notes**, “Introduction to Community Notes,” 2025. Accessed: 2025-02-26.

YouTube, “Community Guidelines,” <https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#detecting-violations> Accessed February 7, 2025.

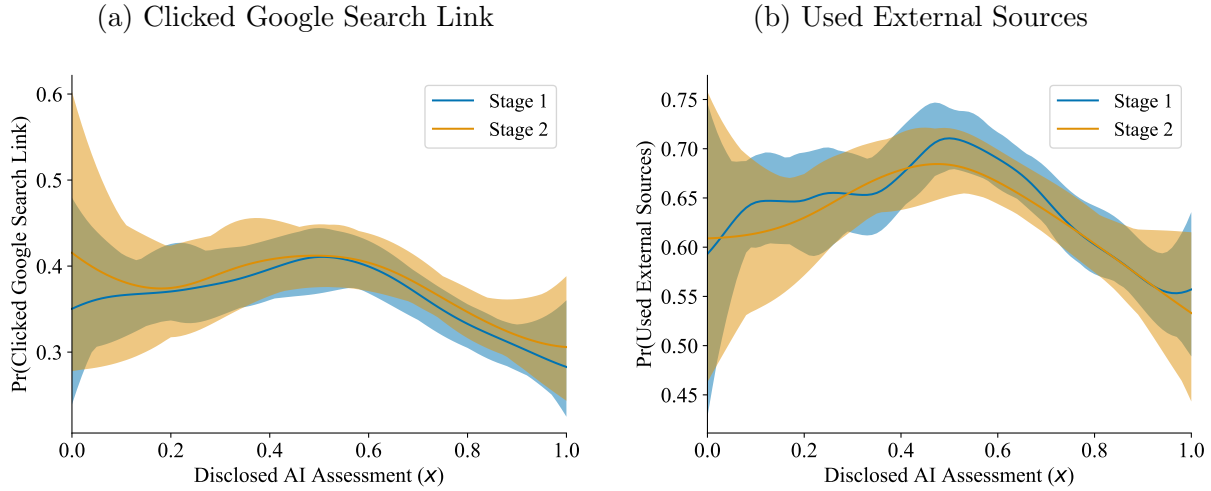
Yu, Feiyang, Alex Moehring, Oishi Banerjee, Tobias Salz, Nikhil Agarwal, and Pranav Rajpurkar, “Heterogeneity and predictors of the effects of AI assistance on radiologists,” *Nature Medicine*, 2024, *30* (3), 837–849.

Appendix

A Data Appendix

A.1 Effort Response for Additional Effort Measures

Figure A.1: Effort Response for Additional Effort Measures



Note: Plots of additional effort measures conditional on x . The curves are estimated via local linear regression and the confidence bands represent bootstrapped 95% uniform confidence bands.

A.2 Balance Tests

All participants in Stage 2 were exposed to all 5 treatments in a random order. To ensure randomization was successful, we test for balance in covariates based on the first treatment encountered. Table A.1 shows the average covariate value by first treatment encountered.

Table A.1: Covariate Balance in Stage 2

	NDNA	FDNA	SL	FDA	NDA	P-value
Total approvals	1112.93	1273.17	1258.97	1016.13	1249.95	0.15
Age	44.31	44.25	44.92	45.20	44.20	0.84
Sex	0.47	0.51	0.54	0.49	0.48	0.31
Share white	0.64	0.63	0.62	0.68	0.65	0.49

Note: Means are computed for each demographic variable conditional on the first treatment seen. “Total approvals” represents the total Prolific studies completed (i.e. approved) by the participant. The p-values are from the joint Wald test that the mean covariates are equal across the five treatments.

Table A.2: Representativeness of Study Participants

	Stage 1		Stagef 2		
	US Census	Sample	P-value	Sample	P-value
Age Distribution					
18-24	0.12	0.12	0.634	0.12	0.281
25-34	0.17	0.18	0.921	0.19	0.071
35-44	0.17	0.17	0.675	0.18	0.149
45-54	0.16	0.16	0.754	0.16	0.630
55+	0.38	0.37	0.334	0.34	0.000
Share Male	0.49	0.49	0.642	0.50	0.561
Share White	0.62	0.63	0.192	0.64	0.010

Note: The means are estimated from Stages 1 and 2. The US Census values are calculated from US Census Bureau population group estimates from 2021 and normalized (accounting for the lack of participants < 18 years of age.) The p-value is computed with the null that the sample average is equal to the US Census value.

* To set up a representative sample, Prolific stratifies the age into five buckets: 18-24, 25-34, 35-44, 45-54 and 55+. Participants are then further stratified based on sex and ethnicity, resulting in a total of 50 subgroups.

Table A.3: Pipeline of Study Participants

Status	Stage 1	Stage 2
Reached Consent	1656	2289
Consented	1648	2279
Began Study	1536	2087
Completed	1501	2000

Note: Table computes the number of participants under various study outcomes. Reached Consent is the number of participants that viewed the consent page. Consented is the number of participants that provided consent. Began Study denotes the number of participants that completed the five practice claims. Completed is the number of participants who successfully completed the study without technical issues.

A.3 Robustness

Table A.4: Average Accuracy by Treatment

Treatment	(1)	(2)	(3)	(4)	(5)
<i>Panel A: No Automation Baseline (β_0)</i>					
Full Disclosure (<i>FDNA</i>)	0.723 (0.004)	0.721 (0.009)	0.723 (0.003)	0.727 (0.005)	0.728 (0.004)
<i>Panel B: Automation Treatment Effects (β_k)</i>					
Full Disclosure (<i>FDA</i>)	0.749 (0.002)	0.752 (0.004)	0.749 (0.002)	0.753 (0.003)	0.754 (0.004)
No Disclosure (<i>NDA</i>)	0.747 (0.001)	0.750 (0.003)	0.747 (0.002)	0.751 (0.003)	0.752 (0.004)
<i>Panel C: No Automation Treatment Effects (β_k)</i>					
No Disclosure (<i>NDNA</i>)	0.689 (0.004)	0.686 (0.008)	0.689 (0.003)	0.693 (0.005)	0.693 (0.004)
Stoplight (<i>SL</i>)	0.725 (0.004)	0.743 (0.008)	0.725 (0.003)	0.729 (0.005)	0.730 (0.004)
Observations	80000	16000	80000	80000	80000

Note: This table summarizes estimates of the average treatment effect on accuracy (proportion correct) in Stage 2 for different specifications. Column (1) estimates the treatment effect without controls or fixed effects. Column (2) only uses data from the first treatment encountered for each participant. Column (3) includes participant and case fixed effects. Column (4) controls for treatment order. Column (5) controls for the number of prior claims encountered. Each model is estimated via OLS. In panel B, the outcomes have been adjusted to account for automation. Standard errors in parentheses are two-way clustered at the participant and claim level.

Table A.5: Average Treatment Effects on Effort (Across)

Treatment	External Sources	Clicked Google	Time Taken (s)
	(1)	(2)	(3)
<i>Panel A: No Automation Baseline (β_0)</i>			
Full Disclosure	0.709 (0.018)	0.457 (0.020)	57.828 (1.722)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
Full Disclosure	-0.428 (0.019)	-0.272 (0.022)	-33.515 (1.864)
No Disclosure	-0.473 (0.019)	-0.306 (0.021)	-38.263 (1.810)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
No Disclosure	0.035 (0.025)	0.041 (0.029)	-0.241 (2.424)
Stoplight	0.003 (0.025)	0.003 (0.028)	0.121 (2.425)
Observations	16000	16000	16000

Note: The average treatment effect is estimated using equation 5. Only the first treatment encountered for each participant is included. This table summarizes the across average treatment effects of different information environments on effort. In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parenthesis.

Table A.6: Average Treatment Effects on Effort (Participant and Case Fixed Effects)

Treatment	External Sources	Clicked Google	Time Taken (s)
	(1)	(2)	(3)
<i>Panel A: No Automation Baseline (β_0)</i>			
Full Disclosure	0.630 (0.004)	0.372 (0.004)	44.551 (0.383)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
Full Disclosure	-0.357 (0.007)	-0.209 (0.007)	-24.515 (0.568)
No Disclosure	-0.412 (0.007)	-0.240 (0.007)	-28.523 (0.595)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
No Disclosure	0.064 (0.006)	0.046 (0.007)	3.749 (0.593)
Stoplight	0.003 (0.005)	0.001 (0.006)	0.091 (0.536)
Observations	80000	80000	80000

Note: The average treatment effect is estimated using equation 5 with additional fixed effects at the participant and case levels. This table summarizes the average treatment effects of different information environments on effort. In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

Table A.7: Average Treatment Effects on Effort (Controlling for Order)

Treatment	External Sources	Clicked Google	Time Taken (s)
	(1)	(2)	(3)
<i>Panel A: No Automation Baseline (β_0)</i>			
Full Disclosure	0.677 (0.009)	0.431 (0.010)	53.340 (0.838)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
Full Disclosure	-0.357 (0.006)	-0.209 (0.006)	-24.531 (0.525)
No Disclosure	-0.412 (0.007)	-0.240 (0.007)	-28.503 (0.566)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
No Disclosure	0.063 (0.006)	0.045 (0.006)	3.689 (0.531)
Stoplight	0.003 (0.005)	0.001 (0.005)	0.008 (0.477)
Observations	80000	80000	80000

Note: Note: The average treatment effect is estimated controlling for treatment order. This table summarizes the average treatment effects of different information environments on effort. In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

Table A.8: Average Treatment Effects on Effort (Controlling for Prior Statements Assessed)

Treatment	External Sources	Clicked Google	Time Taken (s)
	(1)	(2)	(3)
<i>Panel A: No Automation Baseline (β_0)</i>			
Full Disclosure	0.681 (0.004)	0.430 (0.004)	53.473 (0.342)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
Full Disclosure	-0.357 (0.004)	-0.209 (0.004)	-24.523 (0.385)
No Disclosure	-0.412 (0.004)	-0.240 (0.004)	-28.497 (0.385)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>			
No Disclosure	0.063 (0.004)	0.045 (0.004)	3.690 (0.385)
Stoplight	0.003 (0.004)	0.001 (0.004)	0.021 (0.385)
Observations	80000	80000	80000

Note: Note: The average treatment effect is estimated controlling for the number of prior statements assessed. This table summarizes the average treatment effects of different information environments on effort. In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Time Taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parenthesis.

A.3.1 Squared Deviation from Ground Truth

Table A.9: Average Treatment Effects on Accuracy (Deviation from Ground Truth)

Treatment	Correct	Deviation from Ground Truth
	(1)	(2)
<i>Panel A: No Automation Baseline (β_0)</i>		
Full Disclosure	0.723 (0.004)	0.338 (0.003)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>		
Full Disclosure	0.026 (0.004)	-0.006 (0.003)
No Disclosure	0.024 (0.004)	-0.000 (0.003)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>		
No Disclosure	-0.035 (0.005)	0.032 (0.003)
Stoplight	0.002 (0.005)	0.001 (0.003)
Observations	80000	80000

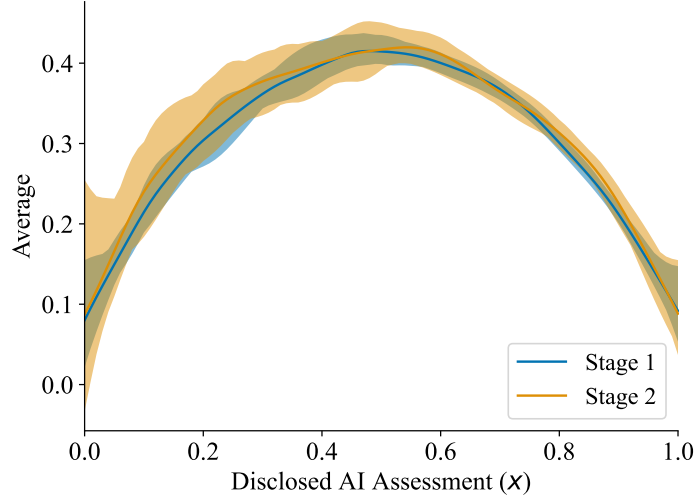
Note: This table summarizes the treatment effects of different information environments on the assessment accuracy as measured by proportion correct (column (1)) and deviation from ground truth (column (2)). In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Standard errors are two-way clustered at the participant and claim level in parenthesis.

Table A.10: Average Treatment Effects on Accuracy (Squared Deviation from Ground Truth)

Treatment	Correct	Mean Squared Error
	(1)	(2)
<i>Panel A: No Automation Baseline (β_0)</i>		
Full Disclosure	0.723 (0.004)	0.206 (0.003)
<i>Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)</i>		
Full Disclosure	0.026 (0.004)	-0.020 (0.003)
No Disclosure	0.024 (0.004)	-0.021 (0.003)
<i>Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)</i>		
NDNA	-0.035 (0.005)	0.021 (0.003)
Stoplight	0.002 (0.005)	0.000 (0.003)
Observations	80000	80000

Note: This table summarizes the average treatment effects of different information environments on the assessment accuracy as measured by proportion correct (column (1)) and deviation from ground truth squared (i.e. mean squared error) (column (2)). In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for automation. Standard errors are two-way clustered at the participant and claim level in parenthesis.

Figure A.2: V Defined Using Deviation from Ground Truth



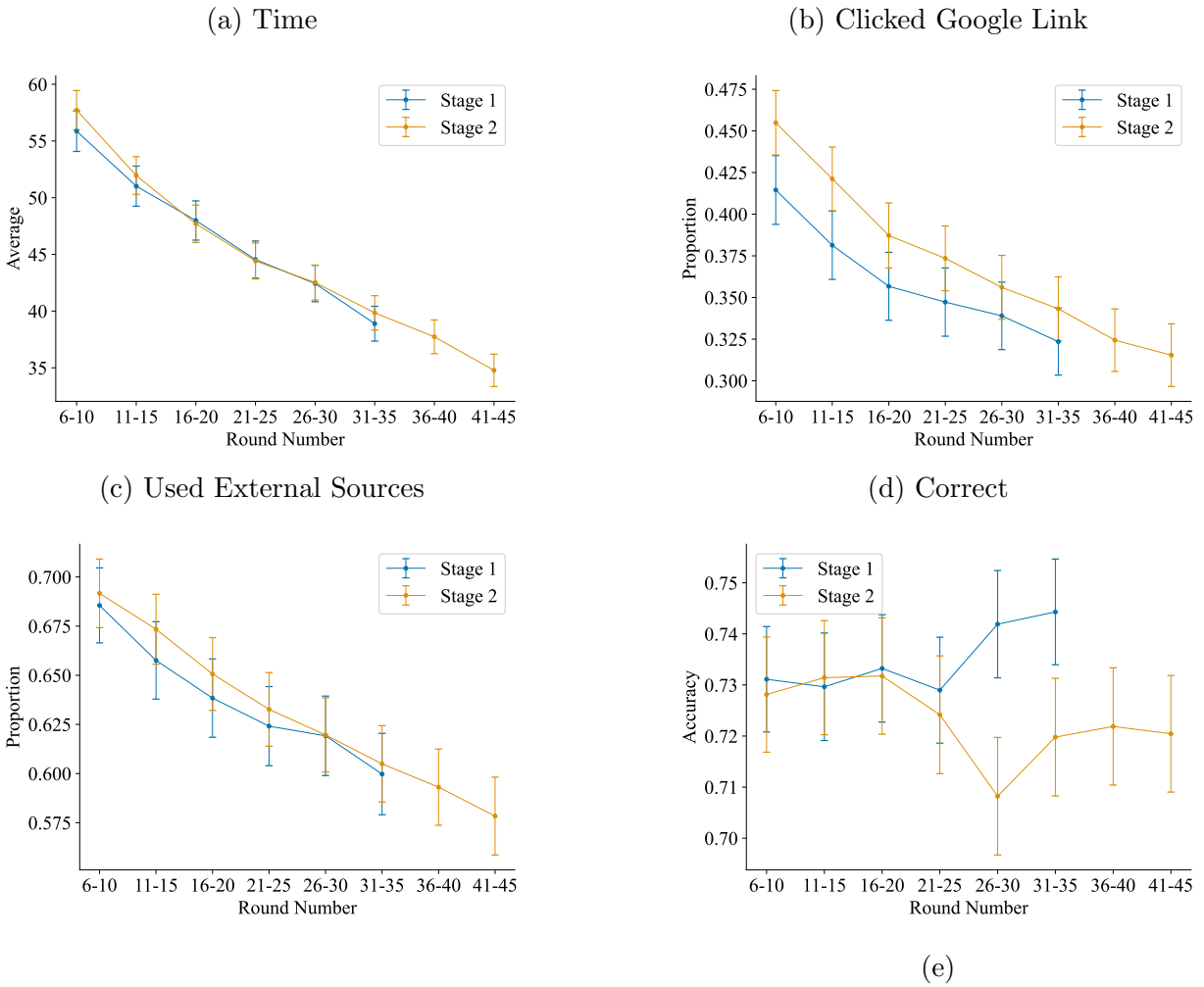
Note: Here $V(\theta)$ is defined as $E[|p_{ij} - \omega_i| | \theta]$. V is estimated using local linear regression from Stage 1 data. The bandwidth is chosen via leave-one-out cross validation to minimize mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level.

A.4 Fatigue and Learning

Participants classified 35 claims in Stage 1 and 45 claims in Stage 2. We test for fatigue and learning effects by estimating the following regression model and plotting β_k in figure A.3.

$$y_{i,j} = \sum_{k \in \text{Intervals}} 1[\text{interval}(i,j) = k] \beta_k + \sum_{k' \in \text{Policies}} 1[\text{policy}(i,j) = k'] \gamma_{k'} + \varepsilon_{ij} \quad (8)$$

Figure A.3: Outcome by Round Number



Note: Figure summarizes outcome by round number. For both stages, data from all treatments is used. The regression model controls for treatment group. Observations from warm up claims are excluded. Claims are grouped into intervals of 5. The 95% pointwise confidence intervals are two-way clustered at the participant and claim level.

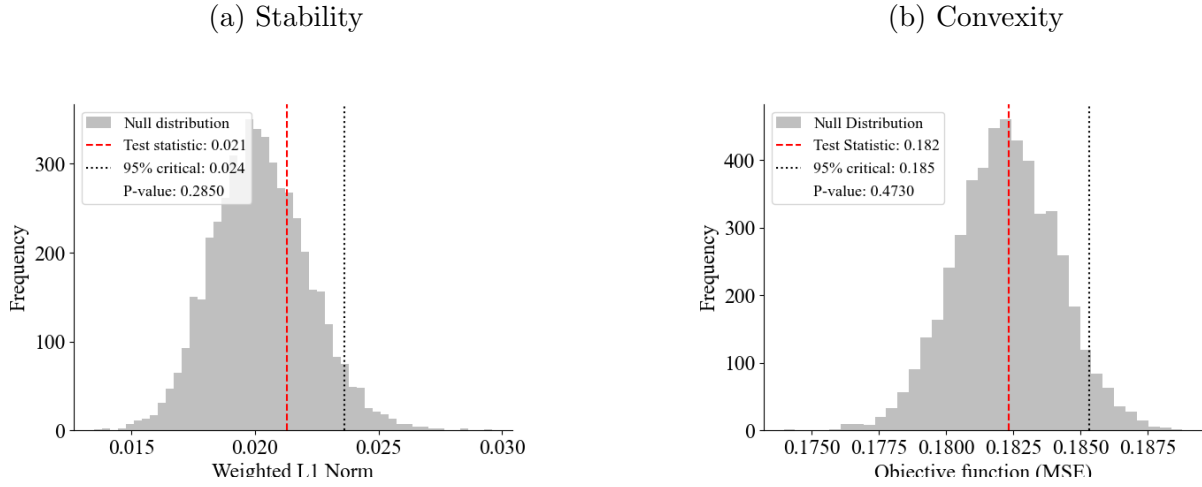
B Stability

Table B.11: Balance: Stage 1 vs Stage 2

	Stage 1		Stage 2		Diff	p-value
	Mean	SD	Mean	SD		
	(1)	(2)	(3)	(4)	(5)	(6)
Correct Classification	0.735	0.441	0.723	0.447	0.012	0.008
Classified as True	0.696	0.460	0.696	0.460	-0.001	0.912
Assessment	0.630	0.329	0.629	0.318	0.001	0.732
Used External Sources	0.637	0.481	0.630	0.483	0.007	0.579
Clicked Google Link	0.360	0.480	0.372	0.483	-0.011	0.383
Time Taken (s)	46.791	43.959	44.551	43.142	2.24	0.032
Observations	45030		16000			
Participants	1501		2000			
Cases per Participant	30		8			

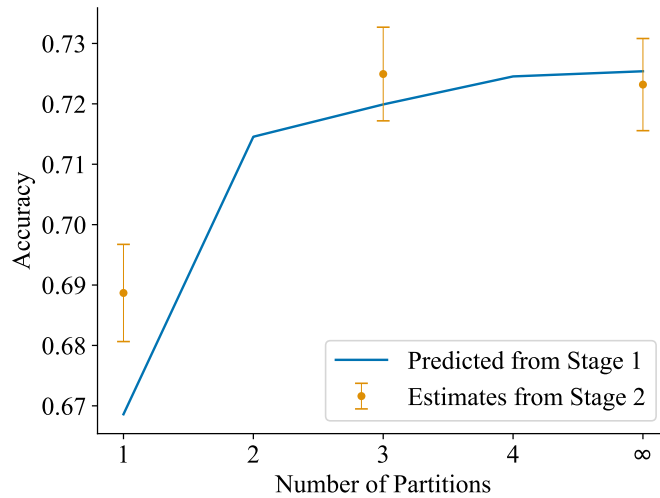
Note: Summary statistics of the experiment using data from the Full Disclosure - No Automation treatment. Columns (1) and (2) present the mean and standard deviation for Stage 1, while Columns (3) and (4) present the same statistics for Stage 2. Column (5) reports the difference between column (1) and column (3), and column (6) reports the p-value that the difference is statistically significant. The p-value in column (6) is from a regression of the outcome on a constant and Stage 2 indicator, with two-way clustering on participants and cases. Correct Classification is an indicator for whether the decision matches the ground truth. Classified as True is an indicator for whether the probability reported > 0.5. Assessment is the probability true reported. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Clicked Google Link is an indicator for whether the participant clicked on the Google link provided by the experimental interface for a particular case. Time Taken (s) is measured in seconds and winsorized to the 95th percentile.

Figure B.4: Test for Stability and Convexity of $V(\theta)$ versus Null Distribution



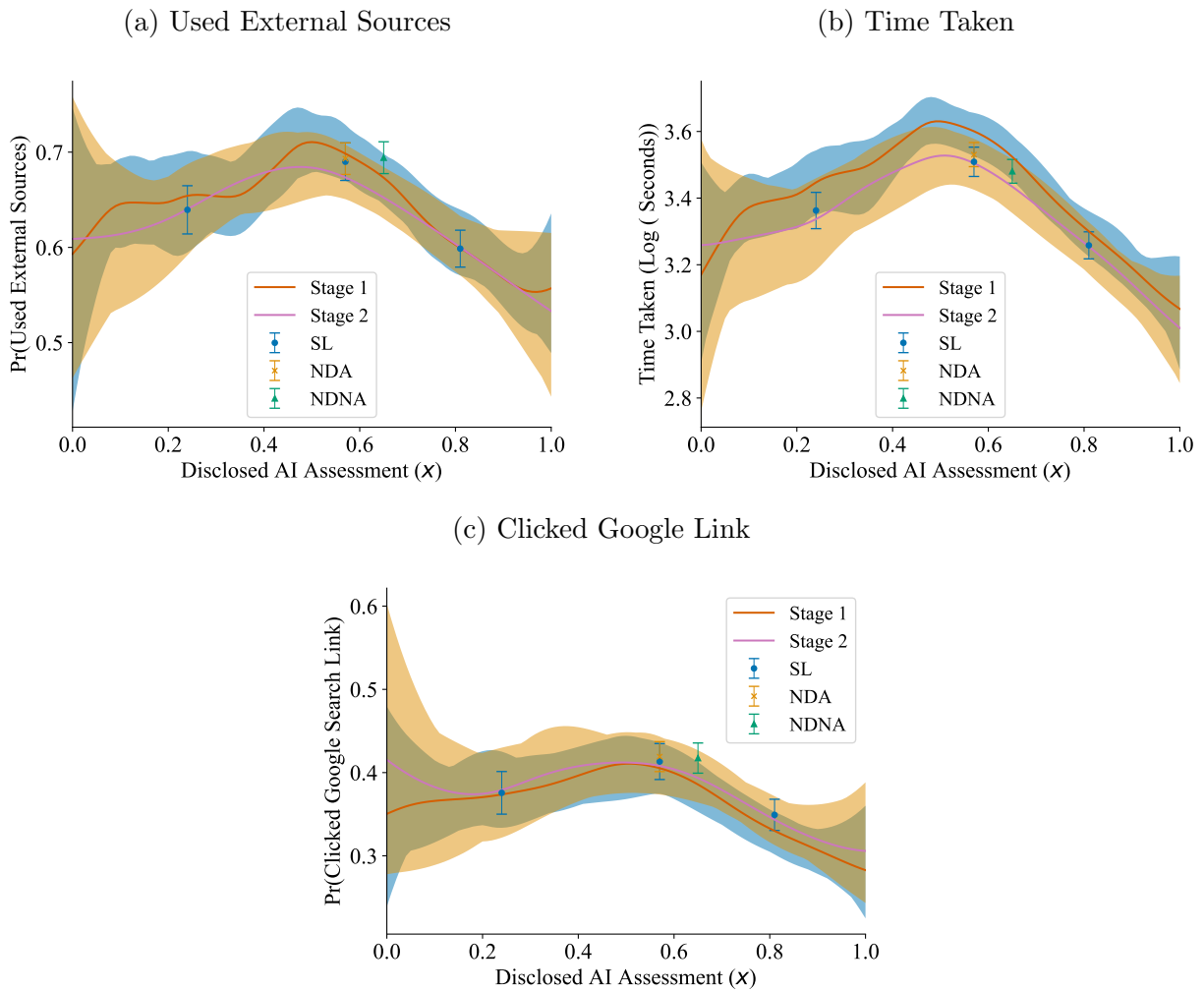
Note: We test for stability and convexity using a two-stage procedure. First, to compute the test statistic for stability, we estimate the weighted L1 norm (weighted by the θ density) between the $V(\theta)$ estimated from Stage 1 and the $V(\theta)$ estimated from Stage 2. Second, we construct a null distribution through bootstrap resampling: in each iteration, we randomly split the Stage 1 data into two groups. Then we compute the unconstrained local linear regression estimate on each half, and calculate the weighted L1 norm between the two unconstrained estimates. The test statistic's percentile in this null distribution provides a p-value for the one-sided test of stability. To compute the test statistic for convexity, we estimate $V(\theta)$ using local linear regression subject to a global convexity constraint, which we implement as a quadratic programming problem, and we save the objective function value. To construct the null distribution, for each bootstrap draw we compute the objective function of the unconstrained kernel regression.

Figure B.5: Stoplight Policy Predicted Accuracy by K



Note: Figure compares the predicted accuracy based on the model with the actual accuracy observed in the experiment. The estimated accuracy from Stage 2 at $K = 1$ is the average accuracy in the No Disclosure + No Automation arm; $K = 3$ corresponds to the average accuracy in the Stoplight + No Automation arm, and $K = \infty$ corresponds to the average accuracy in Full Disclosure + No Automation arm.

Figure B.6: Stability of Effort Response

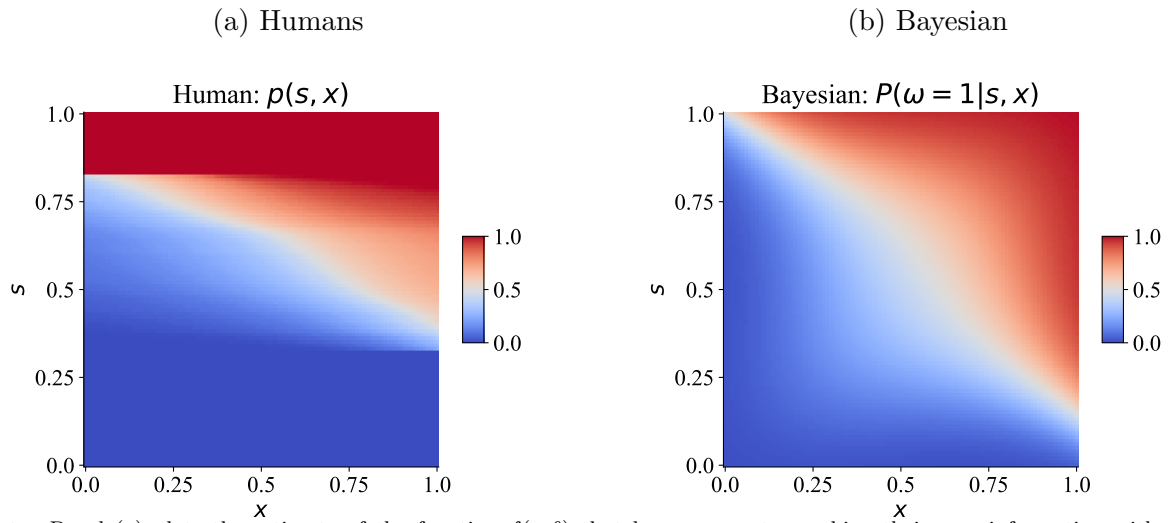


Note: Figure summarizes measures of effort by θ , where the curves are estimated via local linear regression. For both stages, only data from the full disclosure + no automation treatment is used. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Time Spent is the time spent on each case (in seconds). Clicked Google Link is an indicator for whether the participant clicked on the Google link provided by the experimental interface for a particular case. The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level. The average measure of effort by θ by treatment is estimated by regressing the effort outcome on indicators for each AI prediction shown. The 95% confidence intervals are clustered at the participant and case level.

C Mechanisms Appendix

C.1 Additional Empirical Results

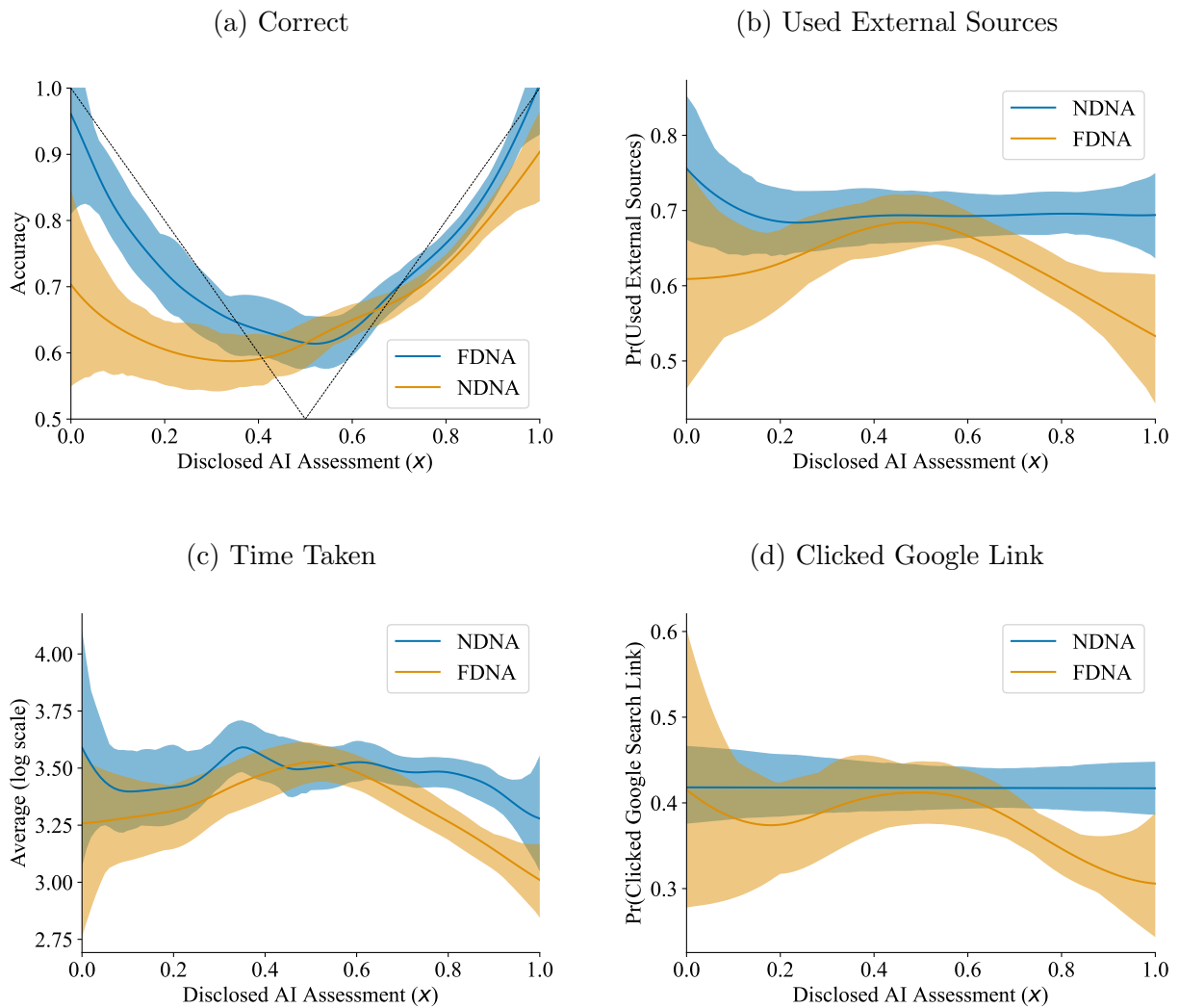
Figure C.7: Human vs Bayesian Update Rule



Note: Panel (a) plots the estimate of the function $f(s, \theta)$ that humans use to combine their own information with the AI assessment. Panel(b) plots the function a Bayesian decision maker uses to combine the two sources of information.

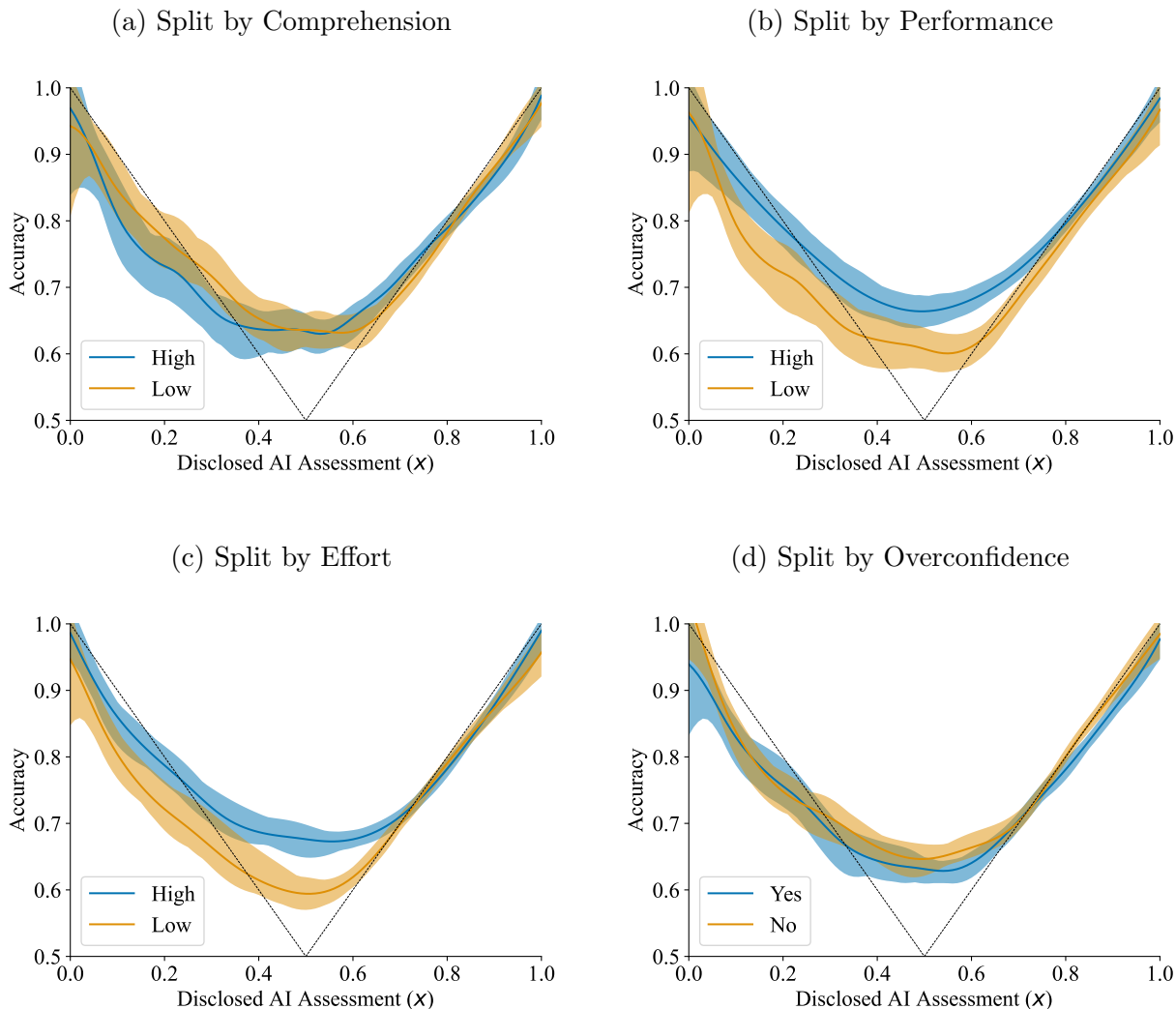
C.2 Heterogeneous Treatment Effects

Figure C.8: Accuracy and Effort by θ



Note: Figure summarizes measures of effort by θ , where the curves are estimated via local linear regression. The figures only use data from Stage 2. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Time Spent is the time spent on each case (in seconds). Clicked Google Link is an indicator for whether the participant clicked on the Google link provided by the experimental interface for a particular case. The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

Figure C.9: Heterogeneity in $V(\theta)$



Note: $V(\theta)$ is estimated using local linear regression from Stage 1 data. $V(\theta)$ is estimated separately for high and low conscientiousness participants, and conscientiousness is measured in four ways: (a) number of comprehension questions answered correctly in the training section (two or less wrong indicates high conscientiousness), (b) performance as measured by a regression of correct minus $\max\{\theta, 1 - \theta\}$ on participant fixed effects, (c) effort as measured by a regression of used external sources indicator on θ , θ^2 , and participant fixed effects, and (d) confidence as measured by a regression of the ground truth on a constant, and the probability reported interacted with participant fixed effects. For figures C.9b, C.9c, and C.9d, the participants are split using the first half of cases encountered, where half the participants are split into the each group, and $V(\theta)$ is estimated on the second half of cases. The bandwidth is chosen via leave-one-out cross validation to minimize mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level. The dashed lines indicate the accuracy of $\max\{\theta, 1 - \theta\}$ that would result under AI automation.

Table C.12: Heterogeneity in Predicted Performance

	SL		FDA		NDA	
	Pooled	Separate	Pooled	Separate	Pooled	Separate
<i>Comprehension</i>						
High	0.750	0.751	0.760	0.762	0.755	0.755
Low	0.715	0.716	0.742	0.743	0.739	0.742
<i>Performance</i>						
High	0.757	0.757	0.765	0.768	0.759	0.761
Low	0.716	0.719	0.740	0.743	0.737	0.742
<i>Effort</i>						
High	0.747	0.747	0.763	0.763	0.757	0.758
Low	0.725	0.727	0.743	0.744	0.740	0.741
<i>Overconfident</i>						
Yes	0.731	0.733	0.751	0.751	0.746	0.747
No	0.750	0.751	0.757	0.758	0.754	0.754

Note: Table displays predicted performance under the three treatments where the pooled policy differs from the separate policy. The pooled column denotes the performance of policies (presented in figure 8) previously estimated on the standard $V(\theta)$ using all the Stage 1 data. The separate column denotes the performance of individually estimated policies for each group (comprehension, performance, effort, and confidence) using the unique $V(\theta)$.

D Alternative Design Approaches

We now discuss alternative approaches that have been proposed in the literature to design Human-AI collaboration. First, we discuss how the sufficient statistic approach differs from the approach taken in the algorithmic triage literature (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023). Second, we discuss an approach that removes the constraint that x represents a calibrated signal and allow the designer to exaggerate the AI signal in an attempt to overcome the under-response to AI that we document above.

D.1 Algorithmic Triage Approach

The algorithmic triage literature focuses on algorithms that selectively automate cases and assign the remaining cases to human decision makers without considering how the human accuracy responds to the automation policy. The sufficient statistic approach has two primary distinctions from the algorithmic triage approach. First, the sufficient statistic approach allows for human beliefs to respond to the designer’s policy. This leads to quantitatively different predictions of accuracy for many automation policies. For example, consider a one-sided automation policy where the designer can only automate True classifications and

assigns the remaining statements to humans with No Disclosure. The optimal one-sided automation policy automates cases where $\theta > 0.58$. We can calculate the predicted performance of this policy as $\gamma^H Pr(\theta \leq 0.58) + E[\theta | \theta \geq 0.58] Pr(\theta > 0.58)$, where γ^H is the predicted performance of humans on cases assigned to them. The sufficient statistic approach predicts $\gamma^H = V(E[\theta | \theta \leq 0.58]) = 65.3\%$ while the algorithmic triage approach treats human performance as fixed and predicts $\gamma^H = E[1[\omega_i = a_{ij}] | \theta \leq 0.58] = 61.2\%$ using data in the No Disclosure + No Automation arm. The difference in performance results from the sufficient statistics approach allowing humans to update their beliefs about the distribution of cases they encounter in response to the automation policy.

A second distinction between the approaches is that the sufficient statistic approach can predict performance for any automation or disclosure policy using only data from Stage 1 (i.e. the data required to estimate $V(x)$). The algorithmic triage approach, however, cannot estimate performance of any policy that involves anything other than either Full Disclosure or No Disclosure (e.g. Stoplight) and requires additional data to predict performance.

D.2 Exaggerating AI Signals to Overcome Automation Neglect

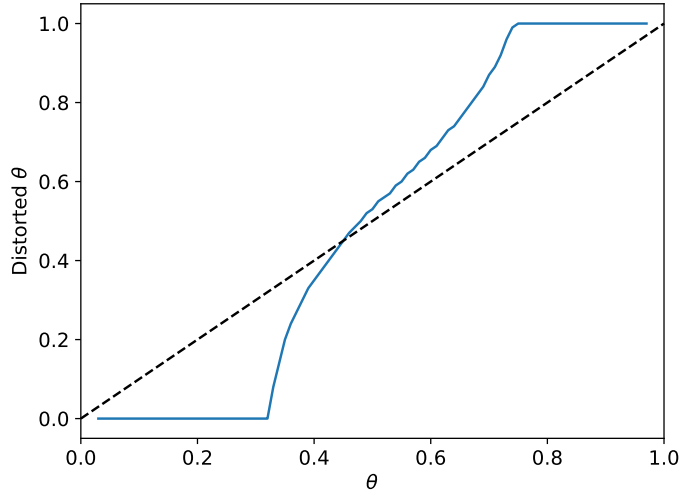
Section 6.2 found that the human participants in our study under-respond to the AI signal relative to a Bayesian decision-maker. This finding is common in the literature on human-AI collaboration (Dietvorst et al., 2015; Agarwal et al., 2023). A natural response to combat such automation neglect is to exaggerate the AI signal (Vodrahalli et al., 2022). That is, the designer can construct a disclosure policy where the AI signal provided to the human is not calibrated. A naïve designer may overestimate the accuracy of such a policy by neglecting to consider how participants update their beliefs when facing a non-calibrated signal. In contrast, our sufficient statistic approach accounts for such updating.

To illustrate this problem in our setting, suppose a naïve designer assumes that the probability that a human decision maker classifies a statement as True is a stable function $T(x, \omega)$ of the disclosed AI assessment x and the ground truth ω , whether or not the assessment is calibrated. Under this assumption, it is optimal for the AI to mis-report any underlying (calibrated) assessment $\theta \in [0, 1]$ as the distorted assessment $\delta(\theta) : [0, 1] \rightarrow [0, 1]$ that maximizes

$$\theta T(\delta(\theta), 1) + (1 - \theta)(1 - T(\delta(\theta), 0)), \quad (9)$$

and the resulting (naïve) expected accuracy is $\mathbb{E}[\theta T(\delta(\theta), 1) + (1 - \theta)(1 - T(\delta(\theta), 0))]$. However, a more plausible assumption is that participants will eventually learn to correctly infer from any reported signal $\tilde{\theta}$ the true conditional probability that $\omega = 1$, $\bar{\delta}(\tilde{\theta}) = \mathbb{E}[\theta | \delta(\theta) = \tilde{\theta}]$,

Figure D.10: Naïve Designer Distortion Map



Note: This figure plots the function $\delta(\theta)$ defined in Equation 9 that maps the actual AI assessment to the distorted AI assessment that a naïve designer would report.

leading to (sophisticated) expected accuracy $\mathbb{E} [V (\bar{\delta}(\theta))]$.

It is straightforward to solve the naïve designer’s problem and compare its naïve and sophisticated expected accuracy. We estimate the function $T(\theta, \omega)$ using a logistic regression with a quadratic term on θ and solve the optimal distortion problem of the naïve designer. Figure D.10 plots the naïve optimal distortion policy $\delta(\theta)$. Due to the AI under-response we have documented throughout the paper, the naïve designer exaggerates the AI signal, for example by reporting $\delta(\theta) = 1$ whenever $\theta \geq 0.75$ and reporting $\delta(\theta) = 0$ whenever $\theta \leq 0.32$.

This naïve optimal policy yields a naïve expected accuracy of 74.7%. This accuracy is very close to that under Full Disclosure + Automation (75.1%). Intuitively, the naïve designer believes that she can nearly replicate automation by exaggerating signals where the AI is confident. However, the sophisticated expected accuracy of this policy is only 73.3%, which is worse than the expected accuracy of 73.5% under Full Disclosure + No Automation. Intuitively, once participants learn and correct the designer’s distortion function, distorting the signal only deprives participants of information (which is sub-optimal since V is convex), rather than correcting automation neglect.

E Estimating Conditional Probabilities

In Section 5.2 and Section 6.2 we non-parametrically estimate a conditional probability of the form $P(\omega_i = 1|W_{ij})$ for a vector of covariates W_{ij} . To do so, we estimate a penalized logistic regression on a polynomial basis expansion of W_{ij} with an elastic-net penalty to avoid overfitting to our data. After the polynomial expansion, we normalize all covariates to be mean zero with unit standard deviation. The elastic-net solves the following optimization problem

$$\max_{\beta} \frac{1}{N} \sum_{ij} (\omega_i \log \hat{p}(W_{ij}, \beta) + (1 - \omega_i) \log (1 - \hat{p}(W_{ij}, \beta))) + \lambda_1 \|\beta\|_1 + \lambda_2 \|\beta\|_2 \quad (10)$$

where $\hat{p}(W_{ij}, \beta) = \frac{\exp(\beta' W_{ij})}{1 + \exp(\beta' W_{ij})}$, $\|\cdot\|_1$ represents the l_1 norm, $\|\cdot\|_2$ represents the l_2 norm, and λ_1, λ_2 are tunable hyperparameters.

We tune the penalty parameters (λ_1 and λ_2) as well as the degree of the polynomial basis expansion to minimize expected out-of-sample loss using grouped 5-fold cross-validation (grouped at the statement level to ensure no data-leakage within a statement) following the recommendations from Hastie et al. (2009) (chapter 7).

Lemmas 1 and 2 in Appendix A of Hirano et al. (2003) shows that the error between a K -th order polynomial approximation of $Pr(\omega = 1|p, \theta)$ and the true function converges to zero, as K increases with the sample size at a specified rate.

F Experimental Instructions

Below are the instructions the subjects received along with the interface-based treatment. These screenshots come from Stage 2. The only differences in Stage 1 are that we estimate the study will take 50 minutes (and adjust the minimum payout accordingly), each individual classifies 35 statements including the practice statements (which changes the maximum possible payouts), and we omit the paragraph “The study will be divided into 5 blocks of 8 statements each. In each block, you will receive assistance from a different AI fact-checker. We will inform you each time you encounter a new AI fact-checker.” from the details of the AI tool.

F.1 Instruction Page 1

Instructions

Welcome! We are a team of researchers from MIT studying collaboration between humans and artificial intelligence (AI) systems.

Your Task

You will be asked to assess whether each of 45 statements is true or false. You may receive information from an AI fact checker to assist you with this task.

We will also provide you with a clickable Google link to the subject of each statement. You are allowed to use the link or any other outside resources.

We expect that this study will take approximately 60 minutes.

Payment

You will earn \$0.35 for each statement that you classify correctly. For example, if you classify all 45 statements correctly, you will earn \$15.75. You will also be eligible for an additional bonus depending on your assessments.

You will receive a minimum of \$8 directly upon completion from Prolific. Any additional payments will be made within two weeks.

Next

F.2 Consent Form

Consent

You have been asked to participate in a study conducted by researchers from the Massachusetts Institute of Technology (M.I.T.). The information below provides a summary of the research. Your participation in this research is voluntary and you can withdraw at any time.

Study procedure: Your main task is to decide whether a statement is true or false. You can use external resources.

Potential risks and benefits: There are no foreseeable risks associated with this study and you will receive no direct benefit (beyond the provided financial incentives) from participating. Your participation in this study is completely voluntary and you are free to choose whether to be in it or not. If you choose to be in this study, you may subsequently withdraw from it at any time without penalty or consequences of any kind. The investigator may withdraw you from this research if circumstances arise.

Privacy & confidentiality

The only people who will know that you are a research subject are members of the research team. No identifiable information about you, or provided by you during the research, will be disclosed to others without your written permission, except if necessary to protect your rights or welfare, or if required by law. In addition, your information may be reviewed by authorized MIT representatives to ensure compliance with MIT policies and procedures. When the results of the research are published or discussed in conferences, no information will be included that would reveal your identity.

Questions

If you have any questions or concerns about the research, please feel free to contact us directly at fact-checking@mit.edu.

Your rights

You are not waiving any legal claims, rights or remedies because of your participation in this research study. If you feel you have been treated unfairly, or you have questions regarding your rights as a research subject, you may contact the Chairman of the Committee on the Use of Humans as Experimental Subjects, M.I.T., Room E25-143B, 77 Massachusetts Ave, Cambridge, MA 02139, phone 1-617-253 6787. I understand the procedures described above. By clicking next below, I am acknowledging my questions have been answered to my satisfaction, and I agree to participate in this study.

Next

F.3 Details of Task

Your Task

You will be asked to assess the likelihood that each of 45 statements is true on a scale from 0% (definitely false) to 100% (definitely true).

If your assessment is greater than 50%, your classification of the statement is "True". If your assessment is less than or equal to 50%, your classification is "False".

We will provide you with a clickable Google link to the subject of each statement. You are allowed to use the link or any other outside resources.

Set of Statements to be Checked

Statements will be randomly selected from a database where approximately 65% of the statements are true and 35% are false.

Next

F.4 Details of AI Tool

Artificial Intelligence (AI) Fact-Checkers

The study will be divided into 5 blocks of 8 statements each. In each block, you will receive assistance from a different AI fact-checker. We will inform you each time you encounter a new AI fact-checker.

Each AI provides its assessment of the likelihood that each statement is true. The AI assessments are correct on average, but not definitive. For example, among all statements that an AI assesses are true with a 70% likelihood, 70% are true, and 30% are false.

Next

F.5 Details of Payment Rule

Payment Rule

You will earn \$0.35 for each statement that you classify correctly. For example, if you classify all 45 statements correctly, you will earn \$15.75.

In addition, you will be entered into a lottery for an additional \$20 bonus, where you are more likely to win the lottery if your assessments are more accurate. If all your assessments are perfectly accurate, your chance of winning the lottery is 10%.

You will receive a minimum of \$8 directly upon completion from Prolific. Any additional payments will be made within two weeks

[Payment Rule Details](#)

[Next](#)

F.6 Comprehension Questions

Comprehension Questions

Before beginning the study, we would like to confirm a few facts through the following comprehension questions. Please answer True or False to the following questions.

Q1: Suppose the AI assigns a likelihood of 40% to a statement. Without reading the statement, what is the likelihood the statement is true?

- 0% 20% 40% 60% Other

Q2: If the AI assigns a 100% likelihood that a statement is true, it could still be false.

- True False

Q3: You are allowed to use outside resources to assist you in this task.

- True False

Q4: How will you be paid for this study?

- An amount depending on the number of correct classifications and the accuracy of your assessments.
 The same amount regardless of your responses in the study.

Q5: Your classification of whether a statement is true or false is the same whether your assessment is 60% or 90%.

- True False

Q6: Your classification of whether a statement is true or false is the same whether your assessment is 45% or 55%.

- True False

Q7: If the statement is False, your chance of winning the lottery is higher when your assessment is 60% than when it is 90%.

- True False

Next