Designing Human-AI Collaboration: A Sufficient-Statistic Approach

Nikhil Agarwal, Alex Moehring, Alexander Wolitzky *

March 18, 2025

WORK IN PROGRESS: DO NOT CITE OR CIRCULATE

Abstract

We develop a sufficient statistic for analyzing counterfactual designs that use AI predictions for either automating decisions or selectively assisting human decision-makers. Although this approach does not fully specify a structural model of human decision-making, it admits models with endogenous effort, biases in belief updating, and endogenous prior beliefs. We deploy and validate the approach in an online experiment on fact-checking, where we design and test optimal AI information disclosure and automation policies. 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 decisions where the AI is confident and delegates decisions while fully disclosing the AI prediction when the AI is uncertain. However, while automation is valuable, the benefit from assisting humans with AI predictions is small.

JEL: C91, D83, D89, D47.

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

*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, Drew Fudenberg, Parag Pathak, Stephen Morris, 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 (Liu et al., 2019; Lai et al., 2021; Mullainathan and Obermeyer, 2022; Kleinberg et al., 2017; Agrawal et al., 2018). Correspondingly, there has been great interest in how AI assistance affects humans' performance (e.g., Noy and Zhang (2023)) and in the design of human-AI collaboration that considers 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 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-dimensional nature of the space of possible collaborative designs, frustrates the search for an optimal design via experimentation.

We develop a sufficient statistic approach (Chetty, 2009) for designing human-AI collaboration for binary classification tasks, where each of several cases must receive a binary classification $a \in \{0, 1\}$. 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 $x \in [0, 1]$.[1] We consider 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 $V(\cdot)$ does not depend on the information disclosure policy, results from the literature on information design (Dworczak and Martini, 2019) imply that $V(\cdot)$ can be used to find the optimal design in this space.

The sufficient statistic approach has important advantages over two natural alternatives. One alternative posits a fully-specified structural model of human behavior and belief updating in order to solve for the optimal designs. This approach is likely to require stronger behavioral assumptions as well as estimates of underlying structural parameters, which may rely on some of the same variation used to estimate our sufficient statistic. A second alternative experimentally tests a range of designs to identify the optimal design. Because the space

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

of potential designs is large, this approach requires conjectures about the optimal design that are likely guided by theory or preliminary quantitative analysis. Moreover, absent sharp theoretical predictions, it is not possible to conclude from this approach that the highest performing tested design is globally optimal.

We implement and validate our approach in an incentivized online experiment on fact-checking in which participants are tasked with checking the veracity of a statement. Fact-checking is an interesting setting to study human-AI collaboration because the veracity of information consumed is an important policy outcome (Lazer et al., 2018). Media organizations, independent fact-checkers and many digital platforms have relied on professional fact checkers (Network, 2023; YouTube, Accessed February 7, 2025; Facebook, Accessed August 12, 2024). The growth in the number of claims to be fact-checked has led to an interest in using both laypeople for fact-checking (Allen et al., 2021; Notes, 2025; Kaplan, 2025), and fully automated solutions (Guo et al., 2022). Professional fact checking organizations are also exploring the use of automated tools in the fact checking process based on generative AI tools (Network, 2023). Understanding how to design AI tools and collaborative systems to improve the productivity of fact-checking is an ongoing challenge.

Besides being an important setting for understanding human-AI collaboration, fact-checking is also convenient for experimental purposes. Fact-checking is easy to explain and can be attempted by inexperienced experimental participants. Measuring accuracy in fact-checking is relatively straightforward, as there are established datasets of true and false statements with curated ground-truth labels: one such dataset is the FEVEROUS dataset (Aly et al., 2021), which we use in our experiment. Finally, fact-checking is representative of other binary classification tasks found in high-stakes settings throughout the economy. For example, binary classification has been used to analyze medical diagnostic decisions (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 $V(\cdot)$ by measuring classification accuracy under varying AI predictions and solves for optimal and approximately optimal designs. We consider both designs where automation is allowed and designs where humans must make the classification decision. This is because in many settings, potentially including fact-checking, there may be a societal preference for humans to make final decisions.[2]. In the second stage, we experimentally implement five designs derived from the first stage estimates and measure the classification accuracy under these designs.

[2]The algorithmic aversion literature, for example, demonstrates that humans often prefer human decision-makers over more accurate algorithmic decisions (Dietvorst et al., 2015; Longoni et al., 2019)

This stage allows us to compare the predictions from our first-stage estimates against the experimental results to validate the sufficient statistic approach.

Our empirical results from the first stage yield several insights and predictions. First, the estimated function $V(\cdot)$ is convex. This 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 (Athey et al., 2020) and empirical results (Dell'Acqua, 2022) that find that partial disclosure of AI information can be optimal because disclosing more precise AI information crowds out human effort in information acquisition. In contrast, 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 AI is confident (x is close to 0 or 1), humans' classification accuracy $V(x)$ is significantly lower than that under automation, $\max\{x, 1-x\}$. This implies that humans under-respond to the AI assessment when updating their beliefs, because just following the AI prediction would increase accuracy when $V(x) < \max\{x, 1-x\}$. It also implies that automating these cases is optimal.[3]

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 providing assistance is approximately optimal. That is, while both humans and AI add value, the value of human-AI *collaboration*—rather than selective automation and delegation—is negligible.[4]

The second stage experiment tests whether the sufficient statistic approach yields accurate predictions for the optimal policies with and without automation derived from the first-stage estimate of $V(\cdot)$. The quantitative results closely match the predictions obtained from the first stage, suggesting that the sufficient statistic assumption is a good guide for designing human-AI collaboration in our context.

We then further analyze the mechanisms that determine the optimal designs and their accuracy. In particular, we decompose the impact of behavioral biases and effort response to AI information.

We first derive a sharp lower bound for the effect of human under-response to AI information on accuracy. The bound is obtained by taking the perspective of a Bayesian observer who has access to both the AI predictions and the humans' reported probability assessments. Here we find that at least 7.7% of incorrect classifications humans make with AI assistance

[3]Under-response to information is a common finding in behavioral economics (Benjamin, 2019). We discuss this literature below.

[4]Our result differs from the finding in Agarwal et al. (2023) that humans assisted with uncertain AI predictions actually perform worse than unassisted humans.

are attributable to errors in belief updating.

We next examine whether humans under-respond to AI information because they are over-confident in the accuracy of their own information or under-confident in the AI. To do so, we estimate the (calibrated) human signals and the update rule our participants use to combine their signals with AI predictions. We find that AI under-response is almost entirely due to over-confidence in own-signal precision: that is, humans' beliefs are too sensitive to their own signals relative to a Bayesian benchmark, but are appropriately sensitive to AI predictions. This contrasts starkly with prior work that attributes AI under-respond to under-confidence in AI signal precision (Agarwal et al., 2023).

Finally, we find that the average impact of effort crowding-out on the precision of human signals is small. However, this impact is large if humans observe a sufficiently confident AI prediction that a statement is false.

Related Literature

The comparison of humans to predictive AI tools has been an active area of research (Mullainathan and Obermeyer, 2022; Kleinberg et al., 2017).[5] Another set of papers compares the accuracy of humans with AI assistance to either the AI alone or humans alone (Angelova et al., 2023; Agarwal et al., 2023; Vaccaro et al., 2024). A strand within this literature is interested in heterogeneous effects by baseline characteristics (Yu et al., 2024). A distinction with this literature is that our focus is on optimally designing the delegation and automation of cases rather than only comparing predictive AI tools with human decisions, either with or without AI assistance.

This 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. These papers abstract away from endogenous changes in human beliefs or effort in response to the set of cases that are delegated or automated. A prior literature argues, based on theoretical (Athey et al., 2020) and empirical results (Dell'Acqua, 2022), that endogenous effort is an important consideration when humans use AI tools. Accounting for endogenous responses on effort and beliefs is an important advantage of our approach.

Our sufficient statistic approach builds on on insights from information design (Kamenica and Gentzkow, 2011). In particular, our sufficient statistic $V(x)$ is the designer's indirect utility from inducing a posterior mean assessment x, following Dworczak and Martini (2019).

[5] A precursor to this literature includes work in the psychology literature comparing human decisions to simple algorithmic decision rules (Meehl, 1954; Dawes, 1979).

To our knowledge, the use of these techniques to design human-AI collaboration is novel. De Clippel and Zhang (2022) study information design to a non-Bayesian receiver. Since our participants are not Bayesian, this problem is related to ours; however, we estimate the designer's indirect utility $V(x)$ and apply standard information design arguments, rather than directly estimating the designer's biased belief updating function (which is the primitive object in De Clippel and Zhang (2022)). Finally, a growing experimental literature on information design focuses on testing the theory's assumptions and predictions rather than using it to design optimal policies and testing their performance (see Fréchette et al. (2022) for examples).

Some of the empirical results parallel findings in prior experiments on biased belief updating. For instance, under-response to new information is a common finding in the behavioral economics literature (see Benjamin (2019), for a survey). Agarwal et al. (2023) also find under-response to AI was also found amongst radiologists, but in contrast to our results, the evidence suggests that professional radiologists neglect AI predictions and are not over-precise. Moreover, our approach for estimating these biases differs in data requirements that we will further discuss below. Relative to this literature, we replicate a standard finding of under-response to external information (the AI signal). However, we go beyond this in decomposing this response into over-confidence in own-signal precision and under-confidence in external information, finding that in our setting under-response is driven almost entirely by the first effect. To do so, we also offer a general definition of over- or under-response to an information source (related to Augenblick et al. (2025)), which may be useful more broadly.

A unique feature of our study is the use of a two-stage experiment in which the first stage estimates a sufficient statistic 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 (Dubé and Misra, 2023; Misra and Nair, 2011; Ostrovsky and Schwarz, 2011). Our approach is closest to that in Ostrovsky and Schwarz (2011), which uses insights from auction theory to derive a sufficient statistic – the distribution of bidder valuations – that is estimated in a first step to solve for the optimal reserve price and test it in a second stage. A qualitative difference from Ostrovsky and Schwarz (2011) is that the dimension of disclosure policies far exceeds the dimension of reserve prices. The sufficient statistics approach avoids the intractable task of experimenting over a large design space.[6] It also avoids estimating a fully specified structural model of behavior, a benefit that has been previously recognized in the context of welfare analysis

[6]Dubé 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.

(Chetty, 2009).[7]

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 given 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. The designer's objective is to maximize the expected accuracy (i.e., the probability of correct classifications) of the overall human-AI collaborative system. The human decision-maker's objective is aligned with the designer's, except that she may also incur unobserved effort costs that are not internalized by the designer, and she may succumb to errors in probabilistic reasoning.

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). For each case $i \in I$, the correct classification (*ground truth*, *state*) is denoted $\omega_i \in \{0,1\}$. For each case, 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 has perfect information about the state then $\theta_i \in \{0,1\}$ with probability 1, while if the AI has no information about the state then $\theta_i = \mathbb{E}[\omega_i]$ (i.e., the ex ante probability that $\omega_i = 1$) with probability 1. 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. If the designer discloses a signal of the AI assessment, the signal induces a posterior probability distribution $\mu \in \Delta([0,1])$ over the AI assessment θ, with mean $x = \mathbb{E}^{\mu}[\theta] \in [0,1]$. In principle, the probability that a human decision-maker correctly classifies a case could depend on the

[7] Misra and Nair (2011) designs an optimal incentive contract based on a structural model of a principal-agent problem, and tests it in a second-stage.

details of the induced distribution μ as well on the underlying policy the AI system uses to disclose signals of its assessment. However, we maintain the following assumption:

Assumption 1 *The probability that a human decision-maker makes the correct classification when they learn that the distribution of the AI assessment is μ is independent of the disclosure policy and depends only on the mean AI assessment $x = \mathbb{E}^{\mu}[\theta]$. We denote the probability of correct classification at mean AI assessment x by $V(x)$.*

The function $V(x)$ is the key sufficient statistic that will allow us to solve for the optimal human-AI collaborative system. Following the information design literature (e.g., Dworczak and Martini (2019)), we refer to the function $V(x)$ as the designer's *indirect utility* from inducing posterior belief x.[8] Under Assumption 1, the indirect utility function $V(x)$ is "structural," in that it is defined independently of the AI disclosure policy.

Assumption 1 implies that there is no benefit to disclosing a non-degenerate probability distribution over AI assessments rather than just the mean assessment. Thus, a disclosure policy can be summarized as a distribution G of mean assessments x. This observation greatly simplifies the formulation of the optimal disclosure/automation design problem.

A leading example where Assumption 1 is satisfied is when decision-makers are Bayesian and obtain a private signal $s \in [0, 1]$ of ω that is independent of θ conditional on ω and x. This example allows the possibility that decision-makers exert costly effort in acquiring their own information about ω, where the precision of their information depends on their effort choice, which in turn can depend on the mean AI assessment x (but does not depend on θ conditional on x). This problem of Bayesian decision-makers making binary classifications on the basis of disclosed signals and conditionally independent private signals has been studied in the information design literature (e.g., Arieli et al. (2023)). In contrast, Assumption 1 is typically violated for Bayesian decision-makers with conditionally dependent private signals.[9]

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 implied mean AI assessment x. For example, this is the case if decision-makers combine their own (conditionally independent) assessment of the state with the mean 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 model of updating in Grether (1980).

[8] Our approach remains valid if the designer's indirect utility $V(x)$ 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 $V(x)$ replacing the probability of correct classification.

[9] In general, for a Bayesian decision-maker, Assumption 1 holds if and only if the (calibrated) 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).

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 subjects are not Bayesian and their information is not conditionally independent of the AI assessment. We take this agreement between predicted and realized accuracy based on Assumption 1 as validation of our sufficient statistic approach.

2.2 The Designer's Problem

We now show how the optimal information disclosure and automation policy can be determined as a function of the indirect utility function $V(x)$. Importantly, an estimate of V circumvents the need for either (i) estimating a structural model of information, updating, and endogenous effort, or (ii) experimenting over a large dimensional space of all possible automation and disclosure policies. In Section 6, we show how to empirically quantify the effect of behavioral biases and endogenous human effort on both the optimal design and the value of the human-AI collaboration.

Under Assumption 1, an information disclosure policy can be summarized by the distribution G of induced mean assessments x. A key result from the information design literature (Blackwell, 1953; Gentzkow and Kamenica, 2016; Kolotilin, 2018) is that such a distribution G is attained by some disclosure policy if and only if it is a *mean-preserving contraction* of the distribution of underlying AI assessments F. Therefore, the maximum expected accuracy attainable by information disclosure alone (when the designer is not permitted to automate the classification) is

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

where $MPC(F)$ denotes the set of all distributions that are mean-preserving contractions of the distribution of AI assessments F.

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

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

The optimal policy is then given by (i) garbling the AI's assessment so that the distribution

of posterior mean assessments x is given by the solution G, (ii) disclosing the realized mean assessment 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 the information provided by the AI (consistent with evidence from prior experiments (Benjamin, 2019) and prior studies on human-AI interactions (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 mean AI assessment x. In our experiment, the distribution of assessments F is given and known. The experiment thus estimates the function $V(x)$. Given this function, we can calculate the optimal mixed disclosure/automation policy and the optimal disclosure-only policy as described above. We can also compare the values of these policies with that of the *full disclosure* policy, where the AI always discloses its assessment, which is given by $\int_0^1 V(x)\, dF(x)$, and that of the *no disclosure* policy, where the AI reveals no information, which is given by $V\left(\int_0^1 x dF(x)\right)$, as well as other policies of interest.

In addition to the 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 mean AI assessment among delegated cases, $\mathbb{E}[\theta \,|\, \theta \notin \Theta^{\text{aut}}]$. The set Θ^{aut} is chosen to solve

$$\max_{\Theta^{\text{aut}} \subset [0,1]} \mathbb{E}\left[\max\{\theta, 1-\theta\} \,\middle|\, \theta \in \Theta^{\text{aut}}\right] \Pr(\theta \in \Theta^{\text{aut}}) + V\left(\mathbb{E}\left[\theta \,\middle|\, \theta \notin \Theta^{\text{aut}}\right]\right) \Pr(\theta \notin \Theta^{\text{aut}}).^{10} \quad (3)$$

[10]This formulation assumes that the designer does not randomize cases with any assessment θ between automation and delegation to human decision-makers. In our setting, the gains from such randomization are trivial.

An important distinction from the prior literature that studies policies that selectively automates cases as a function of the AI assessment (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023, e.g.,) is that we account for how the human decision-maker's beliefs respond to the designer's automation policy. In particular, if the designer automates cases with certain AI assessments θ, we assume that human decision-makers understand this and respond to the correct distribution of θ among the remaining cases that were not automated. In Appendix C.1, we discuss how this response was neglected in prior work, and we quantify the implications of taking it into account.

2.3 Discussion of the Optimal Design

We now describe some implications of the shape of the function V for the optimal disclosure policy and preview our empirical results on the shape of V.

A first observation is that 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.[11]

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 signals 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 subjects learn about ω via a conditionally independent private signal 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.

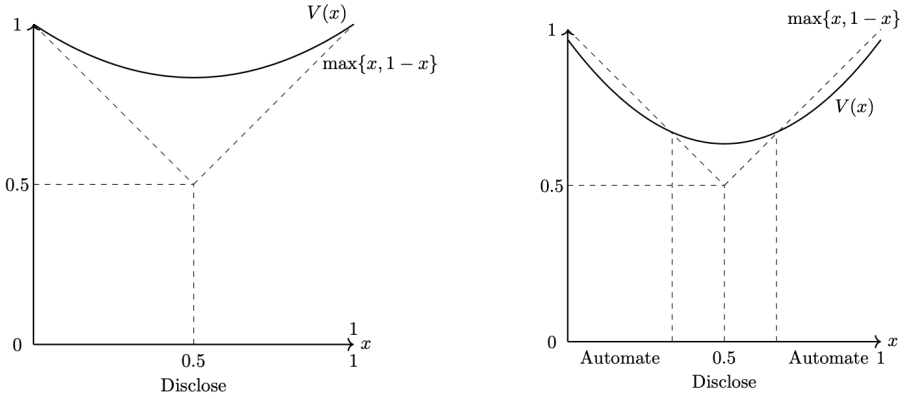
To preview, our empirical results will indicate that V is (approximately) convex in our setting, so fully disclosing the AI assessment is optimal. We also find that there are values of x where $V(x) < \max\{1-x, x\}$, so automation is valuable. Qualitatively, the function V that we estimate has a similar 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 either human effort is sufficiently sensitive to x or if humans exhibit complex behavioral biases. 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. In these settings, it can even be optimal for the designer to delegate some cases where

[11] Intuitively, 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 are 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).

$V(\theta) < \max\{1-\theta, \theta\}$ to the human decision-maker—even though the AI's accuracy on these cases would be higher if it automated them—because pooling them with other cases where $V(\theta) > \max\{1-\theta, \theta\}$ more than compensates for the reduction in accuracy on cases where $V(\theta) < \max\{1-\theta, \theta\}$.[12] Empirically, Dell'Acqua (2022) finds that in his setting human effort is sufficiently sensitive to the disclosed AI signal that overall accuracy is higher with a less precise AI signal, which implies that V is non-convex.

Finally, while we focus on binary classification problems, a similar approach applies for multinomial classification. In the general multinomial case with n possible classifications, the state ω lies in an arbitrary finite set Ω with $|\Omega| = n$. The designer's indirect utility function is now a function $V : \Delta(\Omega) \to \mathbb{R}$ defined on the $n-1$-dimensional simplex (e.g., the probability of correct classification when AI assessment $\theta \in \Delta(\Omega)$ is disclosed to the decision-maker), and the designer's problem is to maximize $\int_{\mu \in \Delta(\Omega)} V(\mu) d\tau(\mu)$ subject to a Bayes plausibility constraint.[13] This is a general Bayesian persuasion problem as formulated by Kamenica and Gentzkow (2011). Thus, while 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, in principle a similar approach applies for multinomial classification problems.

[12] Figure A.1a and A.1b in the appendix illustrate some functions V where such disclosure policies are optimal.

[13] Here, an AI assessment $\theta \in \Delta(\Omega)$ is an $n-1$-dimensional probability distribution, the population distribution of AI assessments is $\phi \in \Delta(\Delta(\Omega))$, and a disclosure policy $\tau \in \Delta(\Delta(\Omega))$ is a probability distribution over distributions $\mu \in \Delta(\Omega)$ satisfying the Bayes-plausibility constraint $\int_{\mu \in \Delta(\Omega)} \mu d\tau(\mu) = \phi$.

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. The first stage estimates the function $V(x)$—the probability of correct classification as a function of the disclosed mean AI assessment $x \in [0, 1]$. The second stage tests the disclosure/automation policies that we find to be optimal under the $V(x)$ function we estimate in the first stage. The two stages are nearly identical except for the AI assistance provided to participants. In the first stage, the AI assessment θ is always disclosed to participants. In the second stage, we test five disclosure/automation policies (including the optimal policies with and without automation), described below. We pre-registered this design and updated the pre-registration to describe the specific policies tested in the second stage as a result of the first-stage estimation.[14] The experiment is implemented on Prolific (`www.prolific.com`). The experimental interface is designed using the o-tree framework (Chen et al., 2016) and accessed through a browser. We next describe the interface and the data that we collect 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 text, 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 2 is 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 $a = 1\,[p > 0.5]$.[15]

In addition to assessments and classifications, we also collect three measures of effort: the time spent on each statement, whether the subject clicked the Google search link, and a participant self-report of whether they used any external source in forming their assessment.

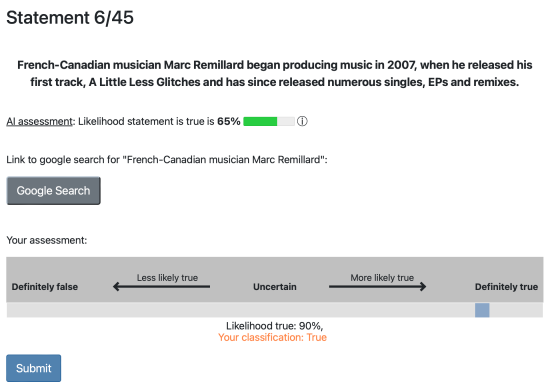
After entering their assessment, participants self-report if they used external sources to form their assessment (Figure 2b). Participants then encounter a feedback screen that includes the AI assessment, the participant's assessment and classification, and the ground

[14]The pre-registration can be found at https://doi.org/10.1257/rct.13990-1.1. We pre-register that we will update the pre-registration after the first stage with the second-stage 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 length of the experiment. Unless otherwise noted, all analyses we present are pre-registered.

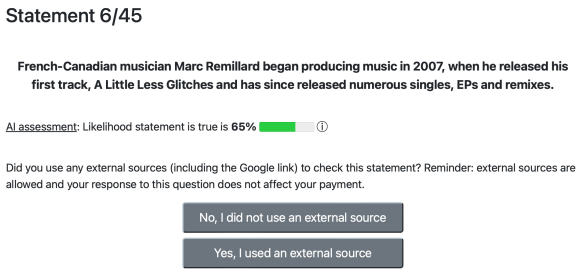
[15]Participants enter their probability assessment through the slider in Figure 2a. The slider button and the text below the slider ("Likelihood true: p%" and "Your classification: a") appear after the participant clicks on the slider to avoid priming.

Figure 2: Screenshots of Experimental Interface

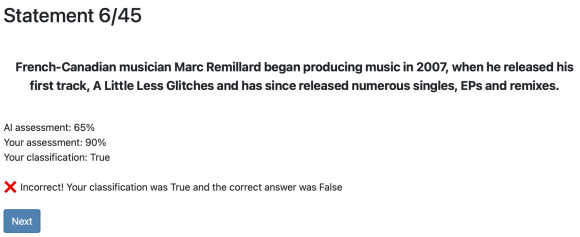
(a) Assessment Screen



(b) Self Reported Effort Screen



(c) Feedback Screen



truth state, True or False (Figure 2c).

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 that no person can participate in both the first and second stages of the experiment. Prolific requires all participants to be at least 18 years of age. We recruit a sample representative of the United States adult population on the dimensions of sex, age, and ethnicity.[16] A summary of demographic information of the study participants is presented in Appendix Table A.2 demonstrating our sample was largely representative of the US adult population. For the first stage of the experiment, we recruited 1,501 participants to assess 30 statements each.[17] We saw minimal attrition, with 97.7% of participants who grant consent and began the study completing stage 1 and 95.8% completing

[16] Certain 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 rounds with non-representative participants.

[17] We pre-registered 1,500 first-stage participants, but one participant who appeared to quit the experiment eventually completed it.

stage 2. The first stage was fielded from July 15, 2024 through August 15, 2024. For the second stage, we recruited 2,000 participants to assess 40 statements each. The second stage was fielded from August 30, 2024 through October 2, 2024.

At the beginning of the experiment, participants receive a high-level overview of the task and the compensation rule they will face. They are then provided with additional information about the task, the interface, and the prior distribution of true and false cases in the database of statements from which we sample. We next introduce the AI fact checker, explaining that it will provide a calibrated assessment of the likelihood that a statement is true. In the second stage of the experiment, we explain that the participants will encounter multiple AI fact-checkers (see Section 5.1). Next, we provide detailed information on the incentives we provide. We explain the incentive scheme 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 incentive scheme. We then test the participant's understanding of the task and automated 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 during this task, and that they understand the incentive scheme. Finally, before beginning the experiment, each participant assesses five practice statements to ensure familiarity with the experimental interface. The full experimental instructions are shown in Appendix D.

Participants are incentivized in two ways to exert effort and provide accurate assessments. The first is a fixed bonus of 35 cents for each statement that they correctly classify. The second is a lottery for an additional $20, where the probability of winning the lottery depends on the accuracy of the subject's probability assessments following Hossain and Okui (2013). Participants are told they will receive a minimum payment if they finish the experiment and that their remaining balance will be paid out within two weeks.[18]

3.3 The Statements

We use the set of statements collected and labeled in Aly et al. (2021), which we refer to as FEVEROUS. The FEVEROUS data set contains approximately 80,000 statements that are labeled as either Supported (True), Refuted (False), or Not Enough Information (NEI). The FEVEROUS statements are constructed by asking annotators to generate statements from a

[18]Specifically, participants receive $7 after completing the first stage or $8 after completing the second stage. This payment was designed to meet the minimum wage for the Prolific platform. It is higher in the second stage where participants assess more cases.

snippet of highlighted Wikipedia text or tables.[19] The statements are politically neutral, so we do not anticipate factors such as motivated reasoning to influence subjects' assessments.

FEVEROUS conducted extensive quality control to ensure the creation of high quality claims and labels. We refer readers to Aly et al. (2021) for full details of this process. In addition to the quality control measures taken in Aly et al. (2021), we remove claims that are not suitable for our study. We first remove the approximately 3% of claims with a label of NEI. We then remove claims with any spelling or grammatical errors flagged by either the rules-based LanguageTool API or GPT-4o.[20] Finally, we remove claims that we determine to be of poor quality, which are mostly claims where the ground truth can change over time, such as for claims that reference an individual's age. In the final set of statements, 65.4% of statements are True.

3.4 The AI Fact-Checker

We use OpenAI's GPT-4o as an automated fact-checker.[21] Specifically, for each claim used in the analysis, we query the OpenAI API with the prompt, "True or False: [claim]" and store the top 20 most likely next tokens along with the probability distribution GPT-4o assess over these possible tokens. We calculate a score θ_i^r for each claim i as follows

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

where token_{ij} is the canonical form (i.e., lower case) of the j^{th} most likely next token and p_{ij} is the probability GPT-4o assigns over the j^{th} token.[22]

While GPT-4o gives us a set of scores s_i for each claim, the model is not calibrated (Achiam et al., 2023). Therefore, we calibrate the model to our data by binning the claims by θ_i^r into 200 bins and calculating the share of claims in each bin that are true. This gives

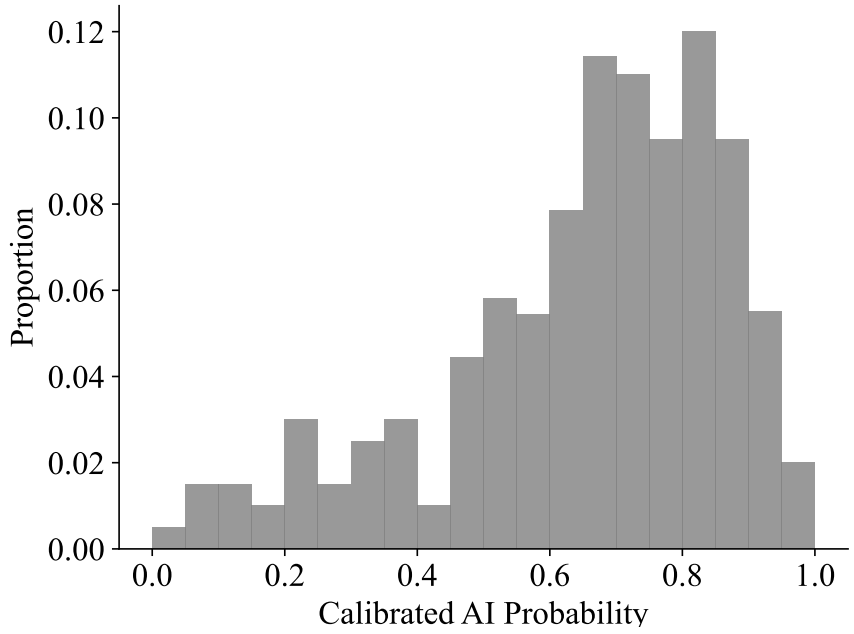
[19] The annotators are different from the individuals used to generate claims. Supported claims require all information within a claim to be verified and supported by evidence. Refuted claims require only a single piece of information within the claim to be refuted by evidence. Claims where not enough information is available on Wikipedia to label the claim as either True or False are labeled Not Enough Information.

[20] For the GPT-4o grammar checking, we queried GPT-4o with the prompt "True or False. The following statement has no grammatical or spelling errors:" followed by each statement. We then discarded statements that GPT-4o assessed to be more likely than not to contain a spelling or grammatical error.

[21] We choose an automated fact-checker based on a large-language model as this fact-checker outperformed tailored automated solutions such as (Aly et al., 2021). Moreover, calibrating the Aly et al. (2021) algorithm resulted in a non-monotonic calibration curve while the GPT-4o based fact-checker had a monotonically increasing calibration curve and the support of the calibrated prediction nearly covered the unit interval.

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

Figure 3: Calibrated AI Assessments



Note: Figure displays the histogram of calibrated AI assessments (from GPT-4o) from the entire population of claims.

us a calibrated AI assessment θ_i for each claim i. We plot the population distribution of θ_i in Figure 3.

4 First Stage Results

In the first stage, we estimate the function $V(x)$ introduced in Section 2.1 and calculate optimal and approximately optimal information disclosure policies with and without automation. We also calculate the predicted treatment effects of the various policies and document how our proxies for effort respond to x.

4.1 Overall Accuracy and Effort

Table 1 describes subjects' accuracy and effort in the first stage of the experiment. 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. However, this similarity masks large heterogeneity in subject accuracy by AI assessment—that is, large variation in $V(x)$—which is key for determining the optimal disclosure/automation policy.

Participants classified 69.6% of cases as true, which exceeds the share of true cases in the database (65.4%). The mean participant assessment of 63.0% is closer to the share of true cases.

Participants appear to exert considerable effort on the task. In 63.7% of cases, participants report utilizing external information sources. In 36.0% of cases, participants clicked the provided link to a Google search for the statement's subject. Finally, on average participants spent 46.8 seconds fact-checking each statement.[23]

Table 1: Summary Statistics, Stage 1

	Phase 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 decision 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) is measured in seconds and winsorized at the 5th and 95th percentiles.

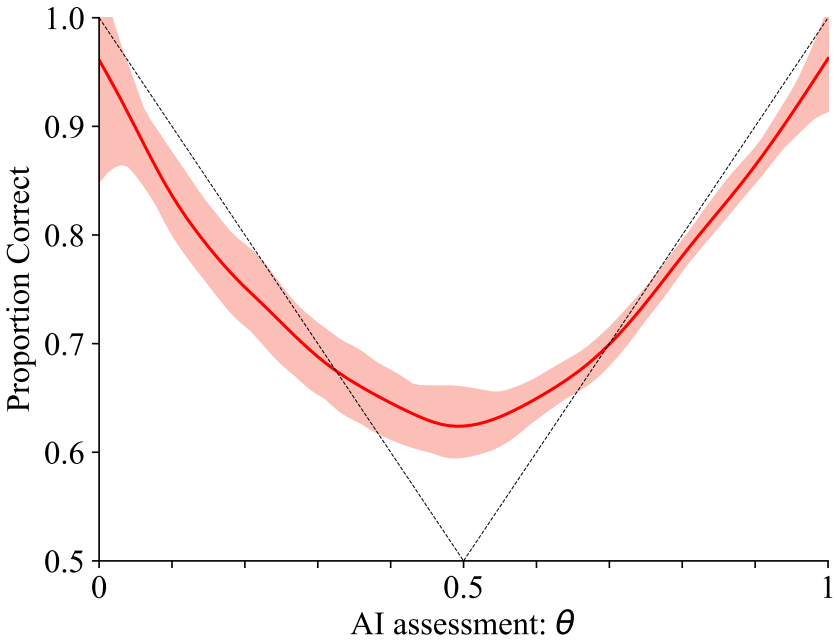
4.2 Accuracy and Effort by AI Assessment θ

Figure 4 presents our estimate of the sufficient statistic $V(x)$ obtained using a local linear regression. Qualitatively, the estimated function $\hat{V}(\theta)$ has a similar shape as the function V in Figure 1b. Two features are key. First, $\hat{V}(\theta)$ appears approximately convex, and a formal test fails to reject that $V(\theta)$ is convex (p =0.466)[24]. Recall that if $V(\theta)$ is convex then full

[23]The median participant in stage 1 took 44 minutes to finish the experiment, including the training, comprehension questions, and the 5 practice statements.

[24]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: $\min_{g,\beta} \sum_i (y_i - g -$

Figure 4: First Stage Estimate of $V(\theta)$



Note: $V(\theta)$ is estimated using local linear regression from the first stage 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 $\max\theta, 1-\theta$ that would result under AI automation.

disclosure of the AI assessment is optimal for all non-automated cases. We thus obtain a key implication for optimal design: **in any optimal disclosure/automation policy, the AI assessment of any non-automated case should be fully disclosed to the human decision-maker**.[25]

Second, on cases where the AI is confident, participants with AI assistance perform significantly worse than they would if they just followed the AI. This is evident in Figure 4, where $V(\theta) < \max\{\theta, 1-\theta\}$ whenever $\theta < 0.33$ or $\theta > 0.69$. Automating fact-checking 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.

$\beta(x_i - t))^2 K_h(x_i - t)$ 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. We save the constrained objective function's value as our test statistic. To construct the null distribution, for each bootstrap draw we compute the objective function of the unconstrained kernel regression. The test statistic's percentile in this null distribution provides a p-value for the one-sided test of convexity. The null distribution and test statistic are plotted in figure A.7.

[25] Appendix A.4 contains an estimate of $V(\theta)$ when the designer's indirect utility is to minimize the deviation of the probability assessment from the ground truth (i.e. $V(\theta) = E\left[|p_{ij} - \omega_i| \mid \theta\right]$).

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-updating bias in a number of experiments on belief updating (see Benjamin, 2019, for a survey) and automation neglect in experiments involving predictive AI assistance (e.g. Agarwal et al., 2023).

While Figure 4 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 *overprecision* in the behavioral economics literature, (see Moore et al., 2015, for a survey). In particular, the function $V(x)$ in Figure 4 can be generated by either a Bayesian with correct beliefs about the precision of their own signal but erroneously low beliefs about the precision of the AI signal, or a Bayesian with correct beliefs about the precision of the AI signal but erroneously high beliefs about the precision of their own signal.

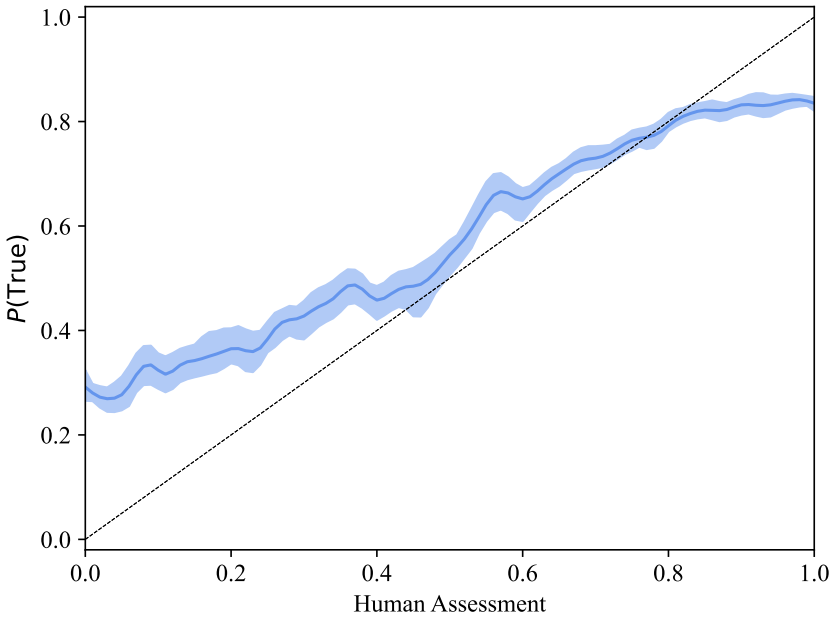
Participants' reported assessments provide evidence of overprecision. Figure 5 plots a calibration curve for stage 1 participants. The slope of the calibration curve is less than 1, indicating overprecision. For example, nearly 30% of statements that participants report are definitely false ($\Pr(True) = 0$) are actually True, and nearly 20% of statements that participants report are definitely True ($\Pr(True) = 1$) are actually False. While Figure 5 provides evidence of overprecision it does not does not provide evidence either for or against automation neglect, and it does not quantify the extent of overprecision. We address these questions in a more structured model of belief updating in Section 6.

We also estimate participant effort as a function of the AI assessment θ. We find that effort decreases as θ moves away from 0.5, the point of maximum uncertainty: see Figure 6. For example, all our effort measures are 10-15 percentage points lower when $\theta = 1$ as compared to $\theta = 0.5$.[26]

We also find evidence that participants' assessments are more accurate when they exert effort. Figure 6b separately estimates accuracy conditional on θ for assessments that took below and above the median response time. We find a positive relationship between response time and accuracy, particularly when the AI is uncertain (i.e. θ is close to 0.5). A similar pattern holds for our other effort measures (see Section A.2). In Section 6, 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.

[26]The reduction in effort when $\theta = 0$ rather than 0.5 is smaller but still substantial.

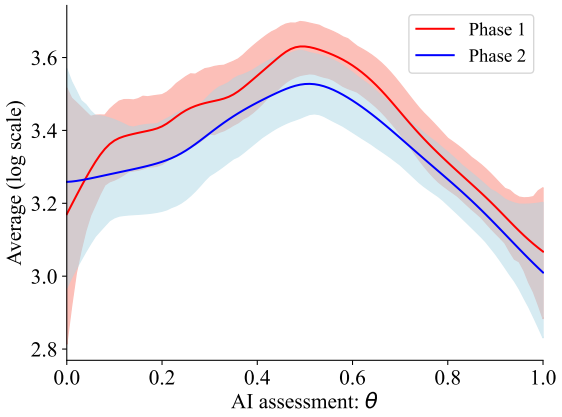
Figure 5: Calibration Curve of Human Assessments



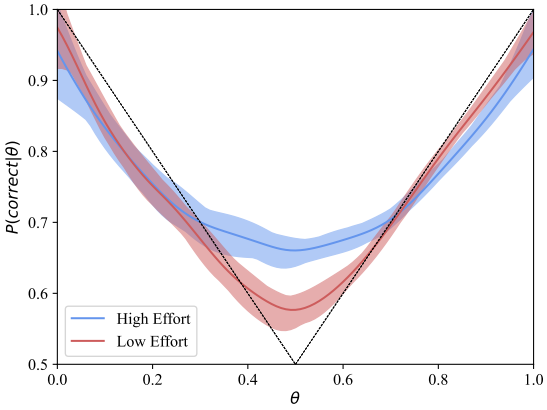
Note: Calibration curve in stage 1. Local-linear regression of ω_i on reported assessments using a Gaussian kernel. Bandwidth is selected to minimize cross-validated mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and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

Figure 6: Effort Response

(a) Time Spent by AI Assessment



(b) Effect of Effort on Accuracy



Note: Figure 6a summarizes the log time taken (in seconds) to assess a statement by θ. The curve is estimated via local linear regression. Figure 6b summarizes the accuracy of participants conditional on effort. The curve labeled "High Effort" is estimated using assessments that took above median time and the "Low Effort" curve is estimated using assessments that took less than the median time. The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

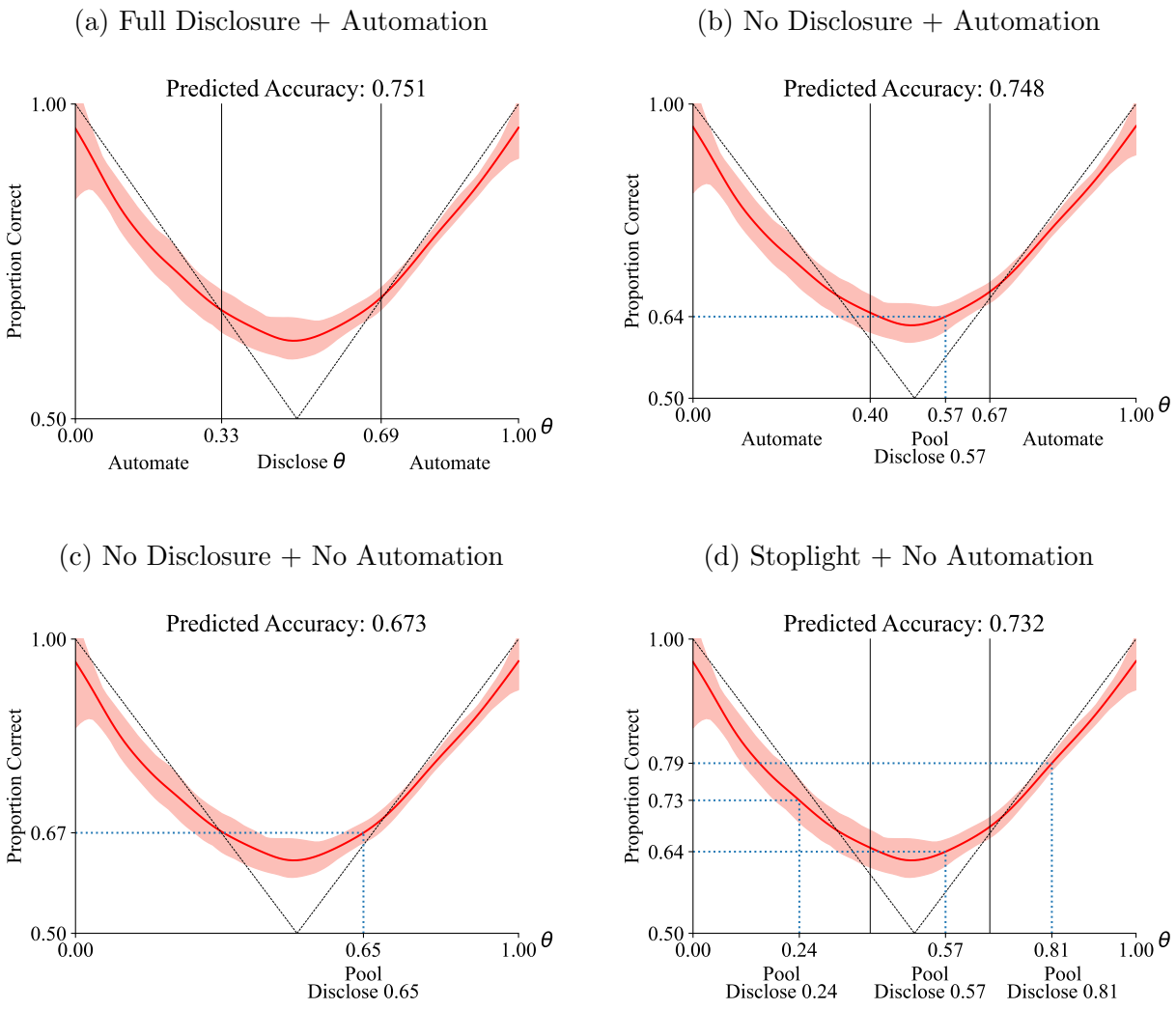
4.3 Optimal and Simple Policies

We now use the estimates of $V(\theta)$ from Figure 4 to solve for the optimal policies, both when automation is feasible and when it is infeasible: that is, we solve the problems (1) and (2) for the estimated function $V(\theta)$. Recall that, since the estimated $V(\theta)$ function is convex, these 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 policy that solves problem (3). Finally, we also consider a simple policy where the AI can only disclose one of three signals. In total, we consider the four policies illustrated in Figure 7, as well as Full Disclosure + No Automation (the optimal policy without automation), which we also ran in stage 1.

The first two policies consider the case where decisions can be automated. Here we compare the optimal policy (Full Disclosure + Automation) and the optimal no-collaboration policy (No Disclosure + Automation).

(a) **Full Disclosure + Automation**: The optimal policy (i.e., the solution to (2)) 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%.

Figure 7: Stage 2 Experiment Overview



Note: Figures summarize the set of policies considered in the experiment. The $V(\theta)$ curve is estimated using local linear regression from the first stage 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 denotes the AI's proportion correct for each value of θ. The dotted lines display the θ disclosed to participants and the associated accuracy predicted by the $V(\theta)$ curve.

(b) **No Disclosure + Automation**: The optimal no-collaboration policy solves problem (3). 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% lower than the optimal policy of Automation + Full Disclosure, the predicted value of direct human-AI collaboration is very small.

(Note that it is optimal to automate cases with a wider range of AI assessments under Automation + No Disclosure than under Automation + Full Disclosure. The key reason why is that a marginal case θ at the boundary of the automation region Θ^{aut} is correctly classified with probability $\max\{\theta, 1-\theta\}$ under automation and with probability $V(\theta)$ under full disclosure, but with probability only $V(\mathbb{E}[\theta|\theta \notin \Theta^{\text{aut}}])$ under no-disclosure. For marginal cases θ we have $V(\theta) > V(\mathbb{E}[\theta|\theta \notin \Theta^{\text{aut}}])$, so automating marginal cases is more beneficial under no disclosure. In addition, the decision to automate or delegate marginal cases affects $\mathbb{E}[\theta|\theta \notin \Theta^{\text{aut}}]$. Since $V(\theta)$ is positively sloped at $\theta = \mathbb{E}[\theta|\theta \notin \Theta^{\text{aut}}]$, 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 0.40—as we move from Full Disclosure to No Disclosure, while the upper threshold of the automation region only slightly decreases from 0.69 to 0.67.)

The remaining policies consider the case when decisions cannot be automated. Here we consider the optimal policy (full disclosure), the no-collaboration policy (no disclosure), and a simple policy that approximates the optimum ("Stoplight").

(c) **Full Disclosure + No Automation**: This is the optimal policy without automation (i.e., the solution to (1)), which is precisely the policy already in place in the stage 1 experiment. Recall that stage 1 accuracy under this policy was 73.5%.

(d) **No Disclosure + No Automation**: 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 + No Automation**: The final policy we consider illustrates the feasibility of approximating full disclosure with a very simple communication policy. Specifically, we calculate the optimal partition of AI assessments into K intervals and disclose the average AI assessment within each interval. For K partitions $[\theta_{k-1}, \theta_k)$ indexed by k,[27]

[27] We set $\theta_0 = 0$ and the last partition to $[\theta_K - 1, \theta_K = 1]$.

we solve the problem

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

Note that $K = 1$ corresponds to No Automation + No Disclosure, and $K = \infty$ corresponds to Full Disclosure + No Automation.

We consider the "Stoplight" policy where $K = 3$ for two reasons. First, predicted accuracy with $K = 3$ is 73.2%, which we expect not to be distinguishable from the predicted accuracy of 73.7% when $K = \infty$.[28] 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 an automated system where 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.[29] 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.[30]

A benefit of the sufficient statistic approach is that it can predict performance for any disclosure or automation policy with only stage 1 data. In particular, the No Disclosure and Stoplight Policies predict human performance for a statement for a counterfactual AI assessment. This feature distinguishes our approach from the algorithmic triage approach for designing human-AI collaboration (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023). Appendix C.1 compares the predicted performance from the two methods, which is quantitatively different for many disclosure policies.

There are two restrictions in our design of the optimal policy that are worth discussing. First, we restrict to disclosure policies in which the AI assessment is calibrated. Appendix C.2 describes policies that allow the designer to exaggerate the AI assessment to overcome the under-response to AI documented above. The benefits of such exaggeration can unravel over time if humans infer that AI assessments are not calibrated.

Second, we do not tailor the policy to predictable heterogeneity across participants. Figure B.12 shows that performance and the sufficient statistic $V(\theta)$ are predictable as a function of

[28] Predicted accuracy for other values of K are shown in Figure A.8.

[29] For 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)

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

baseline comprehension questions, effort, and performance in initial questions. However, table B.12 shows that policies that are tailored to this heterogeneity yield predicted performance similar to that of pooled policies for each group of participants.[31]

5 Second Stage Results

In stage 2 of the experiment, we test each of the above policies—Full Disclosure + Automation, No Disclosure + Automation, Full Disclosure + No Automation, No Disclosure + No Automation, and Stoplight + No Automation—and compare their accuracy to the predictions based on stage 1 data described in Section 4.3. We first describe the experimental protocol—which is the same as stage 1 except for the AI automation/disclosure policy—then describe the treatment effects of the tested policies on accuracy and effort, and finally validate the sufficient statistic approach by comparing the predicted and actual accuracy of the policies we test.

5.1 Experimental Protocol

The experimental protocol is identical to that in stage 1, except participants are randomized into a sequence of the five policies we test. In this stage, each of 2,000 participants assesses a total of 40 randomly selected statements under the five disclosure and automation policies we test. To economize on statistical power, our design includes within-participant comparisons, though we preserve a pure across-participant comparison for robustness. Specifically, each participant assesses eight randomly drawn statements under each of the five disclosure and automation policies. The order of the policies are randomized to ensure that our treatment effect estimates are not confounded by participant learning or fatigue.

5.2 Treatment Effects

We estimate treatment effects for each policy k using the following regression:

$$y_{ij} = \beta_0 + \sum_{k \in \text{ Policies}} 1[\text{ policy } (i,j) = k]\beta_k + \varepsilon_{ij}, \tag{4}$$

where y_{ij} is an outcome for statement i by participant j, the intercept β_0 is the average outcome under the Full Disclosure + No Automation policy, and β_k is the average treatment

[31]The 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.

effect of policy k.[32] Standard errors are two-way clustered at the participant and case level. Table 2 presents the results.

The results are robust to using an across-participant comparison based on the first treatment individuals encounter, to including controls for the treatment order, to including participant fixed effects, and to the number of prior statements assessed by the participant. When using the across-participant design based on the first treatment encountered—Appendix A.5—the treatment effects are approximately similar to our within-participant design, except the baseline for measures of effort is higher. This can be explained by the fatigue effects in Appendix A.5. The estimates from these robustness exercises are presented in Appendix A.5, A.6, A.7, and A.8.

Column 1 reports the key treatment effects on accuracy (share of cases correctly classified).[33] There are four main takeaways.

First, automation is valuable. Automating cases where AI alone outperforms humans with AI assistance increases accuracy by 2.7 percentage points over the baseline of Full Disclosure + No Automation. This effect results from the under-response to AI assessments documented in Figure 4.

Second, both humans and AI add value when automation is feasible. Human accuracy with no AI assistance—i.e., accuracy under No Disclosure + No Automation—is 3.5 percentage points lower than that under Full Disclosure + No Automation and 6.2 percentage points lower than that under Full Disclosure + Automation. At the same time, the accuracy of the AI alone is 73.1%, which is 1.9 percentage points lower than that under Full Disclosure + Automation.

Third, human-AI *collaboration*—disclosing AI assessments to humans, as opposed to just selectively automating cases—has minimal value. The optimal collaborative policy (Full Disclosure + Automation) outperforms the optimal no-collaboration policy (No Disclosure + Automation) by only a statistically insignificant 0.2 percentage points. The intuition for this result is that $V(\theta)$ is rather flat over the optimal range of non-automated cases $\theta \in [0.33, 0.69]$, so there is little gain from information disclosure over this range.

Fourth, simple disclosure policies are approximately optimal. In particular, Stoplight is statistically indistinguishable from Full Disclosure + No Automation.

[32]For treatment arms involving automation, participants only assess cases that are not automated. The estimand of interest is the accuracy and effort of the system, which includes automated cases. We therefore adjust for the automation policy by using the modified outcomes: $y_{ij} \Pr\left(\theta \notin \Theta^{\text{Aut}}\right) + \bar{y} \Pr\left(\theta \in \Theta^{\text{Aut}}\right)$ where Θ^{Aut} is the set of AI assessments being automated in a given policy and $\bar{y}$ is the average outcome among automated statements. For accuracy, $\bar{y} = E\left[\max\{\theta, 1-\theta\} | \theta \in \Theta^{\text{Aut}}\right]$ and $\bar{y} = 0$ for effort measures.

[33]Appendix A.10 displays the treatment effects on the deviation of the probability assessment from the ground truth: $|p_{ij} - \omega_i|$.

Table 2: Average Treatment Effects: Stage 2

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.723	0.630	0.372	44.551
	(0.004)	(0.009)	(0.009)	(0.730)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	0.027	-0.377	-0.221	-26.021
	(0.004)	(0.007)	(0.007)	(0.569)
No Disclosure	0.025	-0.429	-0.250	-29.759
	(0.004)	(0.007)	(0.007)	(0.595)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.035	0.064	0.046	3.749
	(0.005)	(0.006)	(0.006)	(0.586)
Stoplight	0.002	0.003	0.002	0.091
	(0.005)	(0.005)	(0.006)	(0.529)
Observations	80000	80000	80000	80000

Note: The average treatment effect is estimated using equation 4. This table summarizes the average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). In treatments full disclosure + automation and no disclosure + automation, the outcomes have been adjusted to account for the outcomes under automation as described in Footnote 32. Correct is an indicator for whether the decision matches the ground truth. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Clicked Google 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 and winsorized to the 95th percentile.

Figure B.11 plots the heterogeneous treatment effect of Full Disclosure + No Automation relative to No Disclosure + No Automation. Unlike Agarwal et al. (2023), we find only modest evidence that human assessments worsen when presented with an uncertain AI assessment relative to no disclosure.

Columns 2 - 4 report the treatment effects on our three measures of participant effort. Disclosing AI assessments crowds out human effort, consistent with Figure 6. In particular, our effort measures are between 8% and 13% lower under Full Disclosure + No Automation as compared to No Disclosure + No Automation.[34]. However, this effort response is much smaller than that in Dell'Acqua (2022), who found that disclosing more precise AI assessments

[34]Effort under Stoplight is indistinguishable from that under No Automation + Full Disclosure, consistent with these two policies being very similar

reduced effort by nearly 40%. Section 6 presents additional results on the implications of this effort response on the informativeness of humans' signals.

We also investigate heterogeneity in the treatment effect on participant effort by the AI assessment θ. We find that the treatment effect of Full Disclosure relative to No Disclosure on accuracy is largest when the AI is confident, in that θ close to 0 or close to 1 (Figure B.11).

5.3 Validity of the Sufficient Statistic Approach

The key premise of our approach to designing human-AI collaboration is that counterfactual designs can be evaluated on the basis of the estimated function $V(\theta)$ by relying Assumption 1. We assess this premise in two steps. First, we assess the stability of the function $V(\theta)$ under Full Disclosure + No Automation in stages 1 and 2 of the experiment. Figure 8b shows that $V(\theta)$ is stable and we cannot reject $V(\theta)$ estimated in stage 1 equals the $V(\theta)$ we estimate in stage 2 using FD + NA data , although the average accuracy under the Full Disclosure + No Automation arm is lower in stage 2 (72.3%) than in stage 1 (73.5%).[35]

Second, a more demanding requirement is that the accuracy and treatment effects of the policies tested in stage 2 match the predictions based on the stage 1 $V(\theta)$. The quantitative nature of this comparison is significantly more stringent than a qualitative test of the model. We calculate predicted accuracy using both the estimated $V(\theta)$ in stage 1 and an estimate of $V(\theta)$ from stage 2 cases in the Full Disclosure + No Automation arm.

The accuracy of the policies we test in stage 2 closely match the predicted accuracy. Table 3 shows the results. We are unable to reject that the actual accuracy of the Full Disclosure + Automation, No Disclosure + Automation, and the Stoplight policies are equal to the predicted accuracy using either estimate of $V(\theta)$. We are able to reject that the actual accuracy of the Full Disclosure + No Automation condition equals the predicted accuracy from stage 1 and that the actual accuracy of the No Disclosure + No Automation condition equals the predicted accuracy from both stages. We are unable to reject the joint test that stage 2 predicted performances equals the actual accuracy but are able to reject this test using stage 1 predictions.

The difference in the Full Disclosure + No Automation arms stems from the difference in overall performance between the two stages. This may reflect subtle differences in participants between the two rounds or participants learning how to use the AI in stage 1 where

[35]The difference is statistically significant ($p = 0.008$). Participants were also faster in stage 2 than stage 1. This can be explained by participants assessing more claims in stage 2 and their assessments become faster over time (see Appendix A.5).

participants assess 30 statements under FD + NA.

A likely explanation for why predicted accuracy under No Disclosure + No Automation is somewhat lower than actual accuracy is that cases where the AI is confident are also easier for human participants, so that participants' average accuracy under No Disclosure + No Automation is better than their accuracy on cases with the average AI assessment $\mathbb{E}[\theta]$. Figure 8c provides some evidence for this hypothesis. This behavior violates Assumption 1, intuitively due to a failure of conditional independence between the AI's and human participant's information. However, the magnitude of the violation is small: participants' average accuracy under No Disclosure + No Automation is 1.6 percentage points higher than their accuracy on cases with the average AI assessment $\mathbb{E}[\theta]$, 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 73.7%—i.e., the same prediction as under Full Disclosure + No Automation—which exceeds actual accuracy under No Disclosure + No Automation by a much larger 4.8 percentage points.)

Taken together, we find modest evidence that the predicted accuracy differs from the actual accuracy for the Full Disclosure + No Automation and the No Disclosure + No Automation treatment arms. While this may be due to a subtle violation of Assumption 1, it is reassuring that the predicted accuracy of optimal and approximately-optimal policies closely match the actual accuracies we observe. This suggests that despite Assumption 1 not holding exactly, the sufficient statistic approach based on Assumption 1 still provide a useful guide to designing disclosure and automation policies.

For policies that involve no or partial information disclosure, the accuracy of our treatment effect predictions can be visualized in Figure 8a. This figure shows that realized stage 2 accuracy under the disclosed mean AI assessments under No Disclosure + Automation and No Disclosure + No Automation—as well as at each of the three disclosed mean AI assessments under Stoplight—all closely match the corresponding values of $V(x)$ under the stage 1 estimate of the function $V(x)$. Our accuracy predictions are thus on point for each disclosed mean AI assessment, not just on average.

5.4 Bayesian Observer Benchmark

Our finding that $V(x)$ is estimated to be below $\max\{x, 1-x\}$ for some values of x imply that our participants under-respond to the AI assessment. We now bound the impact of this bias on accuracy. This estimate will imply that the optimal policy Full Disclosure + Automation above cannot be significantly improved even if the designer has access to humans' reported

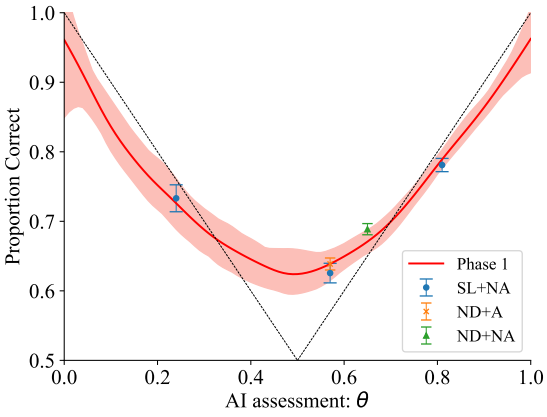
Table 3: Estimated Versus Predicted Accuracy

		Phase 1		**Phase 2**	
Treatment	Estimate	Predicted	P-value	Predicted	P-value
	(1)	(2)	(3)	(4)	(5)
Panel A:					
Full Disclosure + No Automation	0.723	0.737	0.010	-	-
	(0.003)	(0.004)			
Panel B: Automation					
Full Disclosure	0.750	0.751	0.600	0.748	0.556
	(0.002)	(0.002)		(0.003)	
No Disclosure	0.748	0.748	0.968	0.744	0.166
	(0.001)	(0.002)		(0.003)	
Panel C: No Automation					
No Disclosure	0.689	0.673	0.025	0.669	0.033
	(0.004)	(0.006)		(0.009)	
Stoplight	0.725	0.732	0.192	0.720	0.473
	(0.004)	(0.004)		(0.006)	
Joint Test	–	–	0.001	–	0.292

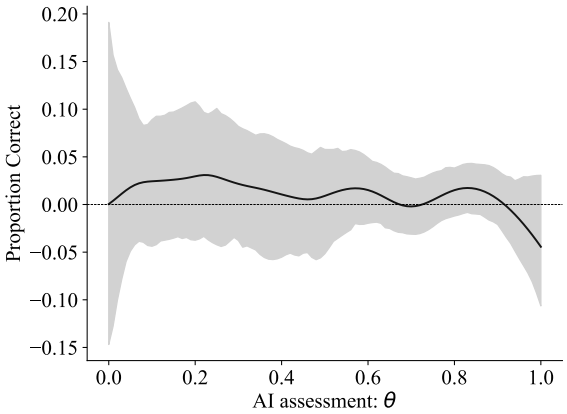
Note: In column (1), Estimate represents the estimated accuracy from the stage 2 data. In column (2), Predicted represents the predicted accuracy computed from the stage 1 data. In column (4), Predicted represents the predicted accuracy computed from the stage 2 data. Columns (3) and (5) contain the P-value from the test where the null sets the Predicted and Estimated values to be the same. Standard errors are in parentheses. Predicted standard errors are computed via block bootstrap, and the Estimate standard errors are two-way clustered at the participant and case level. The stage 2 p-values account for the correlation between the estimates and the predicted accuracy.

Figure 8: Stability

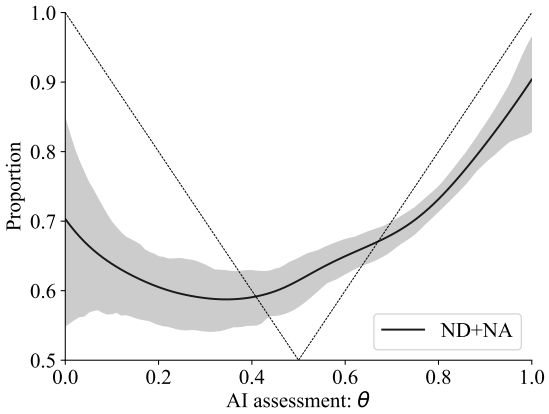
(a) Stability of $V(\theta)$



(b) Difference between Stage 1 and Stage 2 Accuracy



(c) $V(\theta)$ under No Disclosure + No Automation



Note: Figure 8a summarizes proportion correct for stage 1 via the $V(\theta)$ curve estimated by local linear regression. The 45 degree dashed lines denotes the AI's accuracy. The average accuracy by θ by treatment is estimated by regressing a correct indicator on indicators for each AI prediction shown. The 95% point-wise confidence intervals are two-way clustered at the participant and case level. Figure 8b plots the difference in proportion correct between stage 1 and stage 2 by differencing the $V(\theta)$ curves, $V_1(\theta) - V_2(\theta)$, estimated via local linear regression separately for stage 1 and stage 2. The horizontal dotted line is plotted at 0 on the y-axis. Figure 8c summarizes the proportion correct for stage 2 under the No Disclosure + No Automation treatment via the $V(\theta)$ curve estimated by local linear regression. For all three figures, the 95% uniform confidence band is computed via bootstrap accounting for clustering at the participant and case level.

probability assessments p_{ij} and can optimally combine them with AI assessments.

We construct a sharp lower bound for the accuracy of a rational decision-maker with the same information as our participants as a function of the disclosed mean AI assessment x, $V^{\text{Bayes}}(x)$. The bound is obtained by considering an observer with access to the participant's reported assessment p_{ij} and the (disclosed) AI assessment x_i. The observer's accuracy is a lower bound for the accuracy of a rational decision-maker with access to the same information as the participant and the AI assessment x_i, because the decision-maker can reconstruct the participant's report p_{ij} on this basis. Moreover, the bound is sharp in the case where the participant's information is entirely contained in their report p_{ij}.

Formally, we calculate the posterior probability $\Pr(\omega = 1|p, x)$ as a function of the reported participant assessment p and the AI assessment x, and then calculate the accuracy of an observer who classifies a case i with participant assessment p_{ij} and AI assessment x_i as True if and only if $\Pr(\omega_i = 1|p_{ij}, x_i)$ exceeds 0.5. We estimate the function $\Pr(\omega = 1|p, x)$ using a penalized logistic regression and calculate the expected accuracy of this classifier conditional on x to obtain $V^{Bayes}(x)$.[36] In sum, our Bayesian observer benchmark simply takes participants' reports as data and optimally combines it with the AI assessment, thus correcting any mis-weighting of these two types of data.

Figure 9a compares the accuracy of our participants and the optimal policies to those of a Bayesian observer with indirect utility function V^{Bayes}. Recall that Full Disclosure + No Automation is always the optimal policy for a Bayesian with a conditionally independent signal, as observed in Section 2.

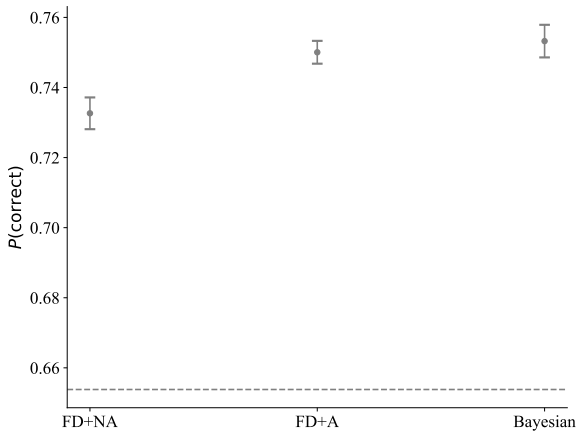
We find that the cost of non-Bayesian updating is substantial – accuracy increases from 73.3% for our participants to 75.3% for our Bayesian observer. We reiterate that this comparison represents a lower bound for the cost of non-Bayesian updating on accuracy. This implies that at least 7.7% of the incorrect classifications humans make under Full Disclosure + No Automation are attributable to deviations from Bayesian updating.

Despite the substantial impact of non-Bayesian updating on accuracy without automation, we find that selective automation can practically eliminate the costs of biased belief updating. In particular, Figure 9a shows that accuracy under Full Disclosure + Automation essentially matches the Bayesian observer's accuracy under Full Disclosure + No Automation. This is a consequence of two properties of the Bayesian observer indirect utility function $V^{\text{Bayes}}(\theta)$, which is illustrated alongside the human indirect utility function $V(\theta)$ in Figure

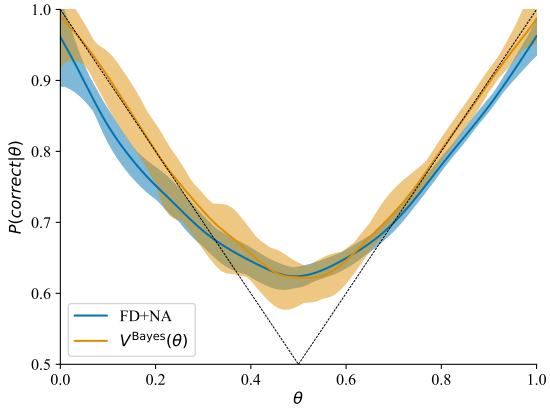
[36] We train a penalized logistic regression with elastic net regularization. We tune the degree of polynomial features and interactions, the regularization parameter, and the $L1$ ratio using grouped five-fold cross-validation (grouped at the statement level) to ensure no data-leakage within a statement. The data is the pooled assessments from stage 1 and stage 2 under Full Disclosure + No Automation.

Figure 9: Comparing Bayesian Decision Maker to Participants

(a) Accuracy of Decisions



(b) Bayesian $V(\theta)$



Notes: Panel (a) plots the accuracy of the participants under the No Disclosure / No Automation (ND+NA), Full Disclosure / No Automation Arm (FD+NA), No Disclosure + Automation (ND+A), and Full Disclosure / Automation (FD+A) treatment arms. Panel (a) also plots the accuracy of fully automating every statement (AI Only) and a lower bound of the accuracy of a Bayesian decision maker in the Full Disclosure / No Automation treatment arm (Bayesian). Panel (b) plots the estimate of $V(\theta)$ for the human participants in stage 1 of the experiment and a lower bound of $V(\theta)$ if humans were Bayesians and optimally combined their own information with that of the AI.

9b. First, $V^{\text{Bayes}}(\theta)$ is nearly indistinguishable from $V(\theta)$ for those AI assessments where delegation to a human decision-maker is optimal (i.e., where $V(\theta) \geq \max\{\theta, 1-\theta\}$). Second, $V^{\text{Bayes}}(\theta)$ is indistinguishable from AI accuracy $\max\{\theta, 1-\theta\}$ for those AI assessments where automation is optimal (i.e., where $V(\theta) < \max\{\theta, 1-\theta\}$). Thus, overall, selective automation achieves the Bayesian observer benchmark.

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 participant biases in belief updating and the impact of effort crowd-out on accuracy.[37] Specifically, we distinguish between participants' *overconfidence* in the precision of their own information and *underconfidence* in the precision of the AI information—which we refer to as *AI neglect*. Distinguishing overconfidence from AI neglect requires an additional assumption, which lets us non-parametrically identify the distribution of participants' private

[37]The analyses in this section were not pre-registered.

information and their model of belief updating. We find that AI under-response is almost entirely due to participants' overconfidence in their own information. This conclusion utilizes a general definition of over- or under-inference that may be useful in other contexts. We also show that the reduction in participant accuracy due to effort crowding-out is fairly modest in magnitude.

6.1 Over- and Under-Inference

We rely on the following general definition of over- or under-inference from a signal. Consider an agent who observes a vector of N real-valued signals $\mathbf{s} = (s_1, \ldots, s_N) \in \mathbb{R}^N$ 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. 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), as then

$$\frac{\Pr(s_n = s|\omega = 1)}{\Pr(s_n = s|\omega = 0)} = \frac{1-p}{p}\frac{\Pr(\omega = 1|s_n = s)}{\Pr(\omega = 0|s_n = s)} = \frac{1-p}{p}\frac{s}{1-s}.$$

Let $f(\mathbf{s}) \in [0, 1]$ denote the agent's assessment of $\Pr(\omega = 1)$ at signal vector $\mathbf{s}$, and let $f^B(\mathbf{s}) = \Pr(\omega = 1|\mathbf{s})$ 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 $\Pr(\omega = 1)/\Pr(\omega = 0)$ from observing a higher signal s_n is always greater than that for a Bayesian: that is, if

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

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

$$\frac{\partial}{\partial s_n}\text{logit } f(s_n, \mathbf{s_{-n}}) > \frac{\partial}{\partial s_n}\text{logit } f^B(s_n, \mathbf{s_{-n}}) \text{ for all } \mathbf{s} \in \mathbb{R}^N \text{ such that } 0 < f(\mathbf{s}) < 1. \tag{5}$$

As far as we know, this definition of over-inference is novel, although it does have some close predecessors. The closest is due to Augenblick et al. (2025), who define a notion of the perceived strength $\hat{\mathbb{S}}(s)$ 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 of Grether (1980) and Agarwal et al. (2023). In particular, if signals are conditionally independent and we rewrite (5) in an equivalent form where the derivatives are taken with respect to logit s_n rather than s_n, then the right-hand side is 1 and the left-hand side is the Grether coefficient on signal s_n.

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 $f(s_{ij}, x_i) = p_{ij}$, which is monotone in s_{ij}.*

Assumption 2.1 imposes two restrictions. First, the distribution of human signals does not directly depend on the disclosure policy or the disclosed AI signal x_i conditional on $\omega_i, e_{ij}, \theta_i$.[38] 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, the signal distribution is the same across human participants conditional on effort.

Assumption 2.2 imposes three restrictions. First, the human assessment p_{ij} is determined only by the human signal s_{ij} and the disclosed AI assessment x_i (and not directly by other observables, such as effort e_{ij}). Second, the assessment is monotone in the human signal.[39] For example, Assumption 2.2 holds if humans are Bayesian with conditionally independent signals (i.e., $s_{ij} \perp \theta_i|\omega_i$). It also holds if humans are quasi-Bayesians who act as if their signals are conditionally independent of the AI signal and may over- or under-weight either

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

[39] It 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.

signal, as in the model of Grether (1980) or the signal-dependence neglect model of (Agarwal et al., 2023). Third, Assumption 2.2 does not allow for heterogeneity in $f(\cdot)$.

We now explain how to identify and estimate the human updating function $f(s, x)$ under Assumptions 2.1 and 2.2.

First, the human signal distribution $G_{\omega_i, e_{ij}, \theta_i}$ is identified from the data under No Disclosure + No Automation. Indeed, in this treatment, the disclosed AI assessment x_i is constant at $\mathbb{E}[\omega]$, while participants report continuous assessment 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, s_{ij}, ω_i, e_{ij}, and θ_i are observable, and hence so is $G_{\omega_i, e_{ij}, \theta_i}$.

Next, let $F_{p_{ij}|\theta_i}$ and $F_{s_{ij}|\theta_i}$ be the conditional CDFs of p_{ij} and s_{ij} given θ_i under Full Disclosure + No Automation. Note that these conditional CDFs are both identified because $F_{p_{ij}|\theta_i}$ is observed under Full Disclosure + No Automation and $F_{s_{ij}|\theta_i}$ can be calculated from $G_{\omega_i, e_{ij}, \theta_i}$ (which is identified from the No Disclosure + No Automation data, as just described) by integrating over the observed variables ω_i and e_{ij}.

Finally, the function $f(s, x)$ can be calculated from the distributions $F_{p|\theta}$ and $F_{s|\theta}$, because, by Assumption 2.2, for any human signal s and AI signal x, we have

$$F_{s|x}(s) = F_{p|x}(f(s, x)),$$

so that inverting the CDF $F_{p|x}$ gives

$$f(s, x) = F_{p|x}^{-1}\left(F_{s|x}(s)\right).$$

Our estimation procedure is analogous to the identification argument. First, we estimate $G_{\omega_i, e_{ij}, \theta_i}$ through a conditional distribution model of $s_{ij}|\omega_i, e_{ij}, \theta_i$ using data in the No Disclosure + No Automation ("ND+NA") arm.[40] Second, we estimate $F_{p_{ij}|\theta_i}$ through a conditional distribution model using data in the Full Disclosure + No Automation ("FD+NA") arm. Third, we estimate the conditional density of $s_{ij}|\theta_i$ in the FD+NA arm by constructing a simulated integral of $G_{\omega_i, e_{ij}, \theta_i}$ over ω_i and e_{ij}. Specifically, we draw 100,000 samples from the joint distribution of s_{ij}, e_{ij}, θ_i, and ω_i from the FD+NA arm. To do this, we estimate the joint distribution of e_{ij}, θ_i, and ω_i in the FD+NA arm using kernel density estimation.[41]

[40] We estimate all conditional CDFs of form $F_{y|x}(z)$ using a logistic regression of the indicator $1[y \leq z]$ on x 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).

[41] In the kernel density estimation, we use a Gaussian kernel for all continuous variables and Silverman's

We then sample from the joint distribution of s_{ij}, e_{ij}, θ_i, and ω_i in the FD+NA arm by first sampling from the joint distribution of e_{ij}, θ_i, and ω_i using an accept/reject sampler and then sampling from the conditional distribution $G_{\omega_i,e_{ij},\theta_i}$ using inverse transform sampling.[42] Finally, we estimate the conditional distribution of $s_{ij}|\theta_i$ using a conditional distribution model (see footnote 40). Given the estimates $\hat{F}_{s_{ij}|\theta_i}$ and $\hat{F}_{p_{ij}|\theta_i}$ we can estimate $f(s,x)$ by inverting $\hat{F}_{p_{ij}|\theta_i=x}$ and calculating $\hat{f}(s,x) = \hat{F}^{-1}_{p|\theta_i=x}\left(\hat{F}_{s|\theta_i=x}(s)\right)$.

6.3 Overconfidence or Automation Neglect?

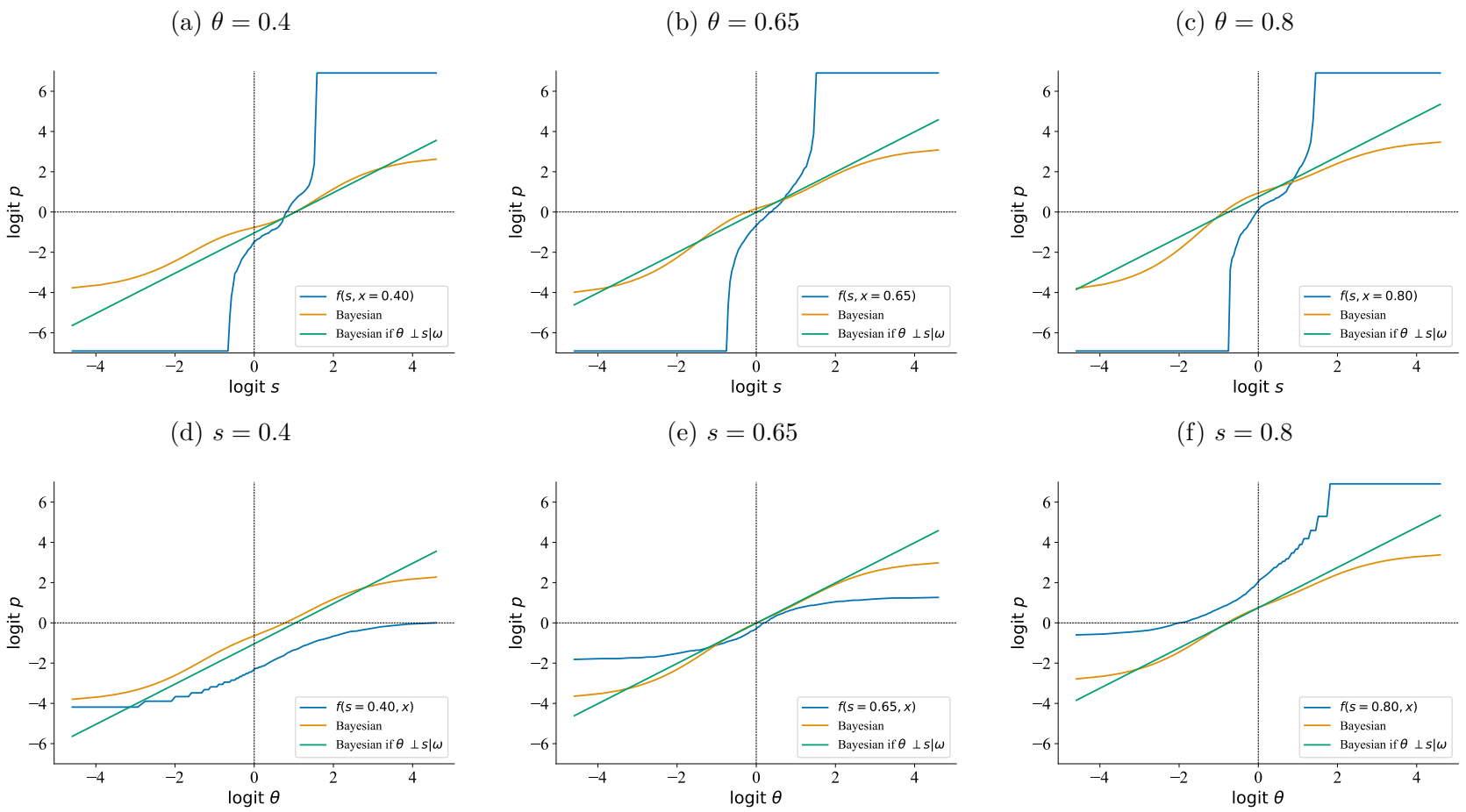
Having estimated our participants' belief updating rule $f(s,x)$, we can now compare it to the Bayesian benchmark and decompose the under-response to AI assessments that we found in Section 4 into overconfidence in one's own signal and AI neglect.

We estimate $f^{\text{Bayes}}(s,x)$ through a penalized logistic regression of ω on s and x in the 100,000 samples of ω, s, and θ from the FD+NA treatment arm described above. We allow for polynomial terms and polynomial interactions and tune the penalty terms as described in footnote 36.

rule to select bandwidths (Silverman, 2018). We manually select a bandwidth of 0 for all binary variables.

[42] In the FD+NA arm $P(\omega_i = 1) = 0.657$, while $P(\omega_i = 1) = 0.649$ in the ND+NA 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.

Figure 10: Human vs Bayesian Update Rule



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

Figure 10 presents estimates of our participants' update function f (blue curve) and the Bayesian benchmark rule f^{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 θ fixed at a specific value while varying the other signal in log-odds space to facilitate interpretation in terms of overconfidence and AI neglect. Appendix Figure B.10 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.

More importantly, Figures 10a–10c show strong evidence of overconfidence: the slope of logit f with respect to s is much larger than the Bayesian benchmark. Such overconfidence results in participants reporting more extreme probability assessments than a calibrated decision maker (as shown in Figure 5) as well as in AI under-response.

In contrast, Figures 10d–10f show weaker evidence of AI neglect: the slopes of logit f and logit f^B with respect to θ are fairly similar, although logit f is somewhat flatter, indicating some degree of AI neglect. Note that the vertical shifts in logit $f(s,\cdot)$ curve relative to logit $f^B(s,\cdot)$ at $s = 0.4$ and $s = 0.8$ reflect overconfidence.

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

$$\text{logit } \tilde{f}(s,x) = \text{logit } f(E[\omega],x) + \text{logit } f^B(s,x) - \text{logit } f^B(E[\omega],x) \tag{6}$$

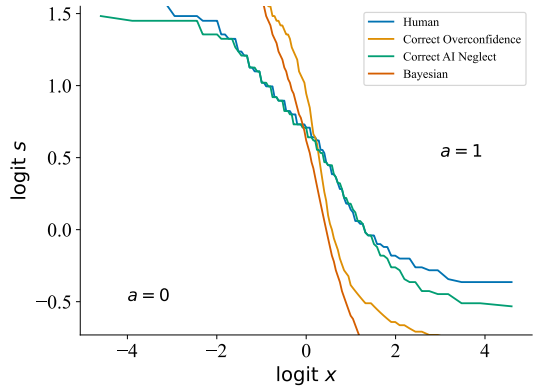
Note that $\frac{\partial}{\partial s}\text{logit } \tilde{f} = \frac{\partial}{\partial s}\text{logit } f^B$, so this correction removes overconfidence. Similarly, we define the human assessment corrected for AI neglect as $\check{f}(s,x)$ such that

$$\text{logit } \check{f}(s,x) = \text{logit } f(s,E[\omega]) + \text{logit } f^B(s,x) - \text{logit } f^B(s,E[\omega]) \tag{7}$$

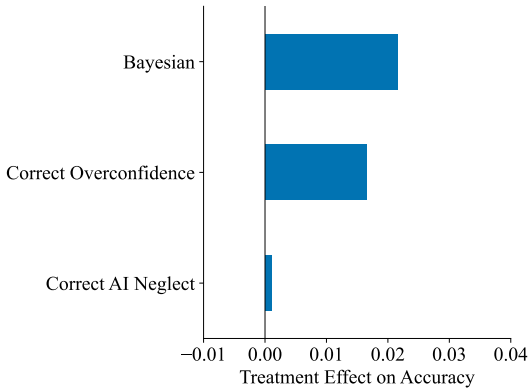
Figure 11a plots the decision threshold in $(\text{logit } x, \text{logit } s)$-space for humans, Bayesians, and humans corrected for overconfidence or AI neglect. We see that the decision threshold for overconfidence-corrected humans is very close to the Bayesian benchmark, while the threshold for AI neglect-corrected humans is very close to uncorrected humans. Correspondingly, Figure 11b 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 in the Bayesian benchmark). These results show that overconfidence—not AI neglect—is the main reason humans deviate from optimal Bayesian decisions.

Figure 11: 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 line corresponds to the set of points where $f(s,x) = 0.5$ for each decision maker. The range of the y-axis corresponds to the support of s. Panel (b) plots the accuracy of each decision maker relative to human participants.

Our result that AI under-response is primarily due to overconfidence rather than automation neglect differs from that in Agarwal et al. (2023), who find evidence for AI neglect but not overconfidence among professional radiologists.[43] One possible hypothesis for this difference is that professional decision-makers (e.g., the radiologists in Agarwal et al. (2023)) understand their own abilities but distrust outside advice, while amateurs (e.g., our participants) understand their own abilities less well but are more open to advice.

6.4 Impact of Effort Crowd-Out on Human Signal Quality

Our identification of human signals s also lets us measure the impact of effort crowd-out on the precision of human signals. Specifically, we use estimates based on Assumptions 2.1 and 2.2 to compare the quality of the human signal s under Full Disclosure + No Automation and No Disclosure + No Automation 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 measures of effort and human signal precision.

Panel A shows that disclosing the AI assessment reduces all of our 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 consistent with the overall treatment effects on effort documented in Section 5.

Panel B shows that this effort crowding-out also reduces human signal precision and

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

Table 4: Impact of Disclosing AI Assessment on Effort and Quality of AI Assessment

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

Note: This table summarizes the impact of reduced effort response in the Full Disclosure + No Automation condition (FD + NA) relative to the No Disclosure + No Automation condition (ND + NA) on the informativeness of the human signal s. We report all measures averaging over all statements as well as conditional on the AI assessment (θ). Panel (a) summarizes the difference in effort exerted by participants in the FD + NA arm relative ot the ND + NA arm. Panel (b) reports the treatment effect of FD + NA relative to ND + NA on the accuracy of the human signal ($P(\text{Correct} = P(1[P(\omega = 1|s) > 1/2] = \omega)$) and the probability of classifying a statement as True $P(\text{True}) = P(P(\omega = 1|s) > 1/2)$.

increases the probability that the human signal is insufficient to overturn to the prior favoring classifying cases as True.[44] The reduction in precision and increase in the probability of classifying cases as True is concentrated on statements that the AI is confident are false ($\theta < 0.25$).

Overall, Table 4 shows that effort crowding-out due to AI disclosure reduces human signal precision and contributes to the value of selective automation, especially for statements that the AI is confident are false.

7 Conclusion

Although the space of potential human-AI collaboration designs is large, we show that an optimal design can be recovered using estimates of a sufficient statistic. This approach can account for behavioral biases and endogenous responses to the design, aspects that have been previously ignored in the "algorithmic triage problem." Our experiments on fact-checking validate this approach, suggest that behavioral biases are an important consideration in the designer's problem, and show that simple disclosure or simple automation policies are approximately optimal.

In addition to the optimal design of human-AI collaboration, our sufficient statistic can also be used for evaluating changes in the quality of the AI since an alternative predictive

[44]Recall that 65.4% of cases in the database are true.

AI tool would generate the distribution $\tilde{F}$ over the assessments θ instead of F. In particular, this value can be calculated under various forms of human-AI collaboration. We leave this avenue of research for future work.

Our experiment is designed to focus on the design of human-AI collaboration, and we consequently do not investigate learning, fatigue, or heterogeneity by participants. During the short duration of our experiment, we did not observe noticeable changes in accuracy of our participants. Likewise, we also did not find substantial predictable heterogeneity in treatment effects.

A promising avenue for future research is to broaden the scope of human-AI collaboration design to move beyond binary classification and prediction problems. While this class of problems includes many important examples, extending the analysis to higher-dimensional prediction problems is valuable. Designing human-AI collaboration for problems other than prediction problems is also an open area for research.

The space of policies considered can also be enlarged. While we find evidence of an effort response when AI information is disclosed, we do not consider the joint design of an incentive contract and human-AI collaboration.

Finally, biases in belief updating have now been documented in several examples of human-AI collaboration. Approaches for counteracting these biases would be valuable. One approach would involve exaggerating the AI assessments to mitigate under-response bias. Studying such a policy would require an understanding of how humans update their use of AI assessment when they are misled. Another approach is to train humans how to better utilize AI predictions. Answers to these questions would be valuable to the extent that humans can access information that is not available to AI tools.

References

Achiam, Josh, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat et al., "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

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.

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.

Dawes, Robyn M, "The robust beauty of improper linear models in decision making.," *American psychologist*, 1979, *34* (7), 571.

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.

Dworczak, 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.

Fréchette, 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, "Competition in persuasion," *The Review of Economic Studies*, 2016, *84* (1), 300–322. https://doi.org/10.1093/restud/rdw052.

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.

Hossain, Tanjim and Ryo Okui, "The binarized scoring rule," *Review of Economic Studies*, 2013, *80* (3), 984–1001.

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.

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.

Liu, Xiaoxuan, Livia Faes, Aditya U Kale, Siegfried K Wagner, Dun Jack Fu, Alice Bruynseels, Thushika Mahendiran, Gabriella Moraes, Mohith Shamdas, Christoph Kern, Joseph R Ledsam, Martin K Schmid, Konstantinos Balaskas, Eric J Topol, Lucas M Bachmann, Pearse A Keane, and Alastair K Denniston, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis," *The Lancet Digital Health*, October 2019, *1* (6), e271–e297.

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.

Meehl, Paul E, *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence.*, University of Minnesota Press, 1954.

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.

Moore, Don A, Elizabeth R Tenney, and Uriel Haran, "Overprecision in judgment," *The Wiley Blackwell handbook of judgment and decision making*, 2015, *2*, 182–209.

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.

Network, International Fact-Checking, "State of Fact-Checkers 2023," Technical Report, Poynter Institute 2023.

Notes, X Community, "Introduction to Community Notes," 2025. Accessed: 2025-02-26.

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," in "Proceedings of the 12th ACM conference on Electronic commerce" 2011, pp. 59–60.

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.

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.

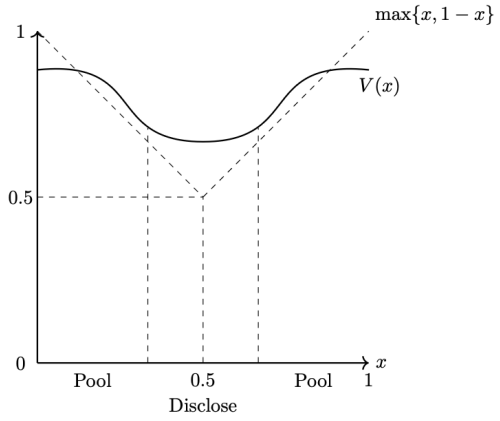
Appendix

A Data Appendix

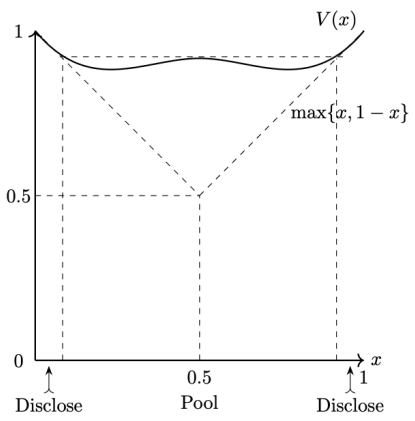
A.1 Optimal Design

Figure A.1

(a) Conceal Extreme Signals



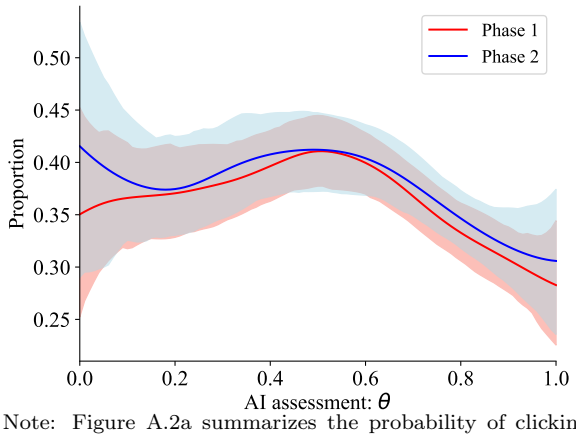
(b) Conceal Moderate Signals



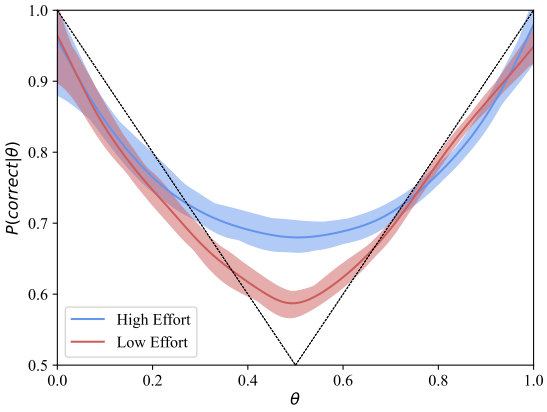
A.2 Effort Response for Additional Effort Measures

Figure A.2: Effort Response - Clicked Google Button

(a) Clicked Google by AI Assessment



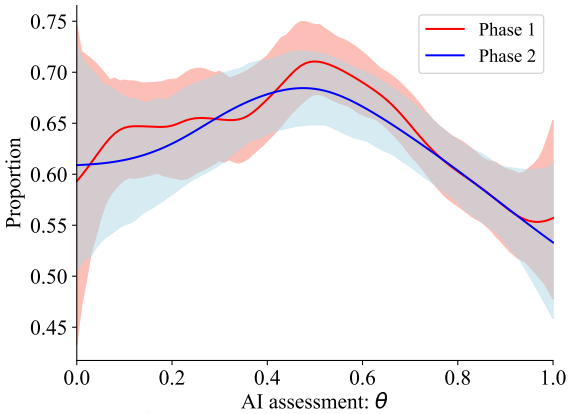
(b) Effect of Effort on Accuracy



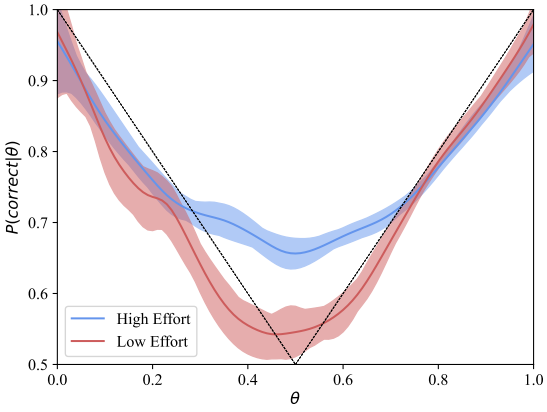
Note: Figure A.2a summarizes the probability of clicking the google button by θ. The curve is estimated via local linear regression. Figure A.2b summarizes the accuracy of participants conditional on clicking the google button ("High Effort") or not ("Low Effort"). The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

Figure A.3: Effort Response - Used External Sources

(a) Used External Sources by AI Assessment



(b) Effect of Effort on Accuracy



Note: Figure A.3a summarizes the probability of self-reporting using external sources by θ. The curve is estimated via local linear regression. Figure A.3b summarizes the accuracy of participants conditional on reporting using external sources ("High Effort") or not ("Low Effort"). The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

A.3 Balance Tests

All participants in phase 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 Phase 2

	ND+NA	FD+NA	SL+NA	FD+A	ND+A	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

		Phase 1		**Phase 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 the first and second phases of the experiment. 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	**Phase 1**	**Phase 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.4 Unadjusted treatment effects

Table A.4: Average Treatment Effects: Phase 2 (Unadjusted)

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.723	0.630	0.372	44.551
	(0.004)	(0.009)	(0.009)	(0.730)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	-0.076	0.054	0.036	5.563
	(0.005)	(0.006)	(0.006)	(0.570)
No Disclosure	-0.085	0.063	0.047	6.325
	(0.005)	(0.006)	(0.006)	(0.560)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.035	0.064	0.046	3.749
	(0.005)	(0.006)	(0.006)	(0.586)
Stoplight	0.002	0.003	0.002	0.091
	(0.005)	(0.005)	(0.006)	(0.529)
Observations	80000	80000	80000	80000

Note: The average treatment effect is estimated using equation 4. This table summarizes the average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). In treatments full disclosure + automation and no disclosure + automation, the outcomes have not 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.5 Robustness

Table A.5: Average Treatment Effects: Phase 2 (Across)

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.721	0.709	0.457	57.828
	(0.009)	(0.018)	(0.020)	(1.722)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	0.031	-0.449	-0.285	-35.342
	(0.010)	(0.019)	(0.021)	(1.844)
No Disclosure	0.029	-0.491	-0.317	-39.772
	(0.009)	(0.018)	(0.021)	(1.797)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.035	0.035	0.041	-0.241
	(0.012)	(0.025)	(0.029)	(2.424)
Stoplight	0.022	0.003	0.003	0.121
	(0.012)	(0.025)	(0.028)	(2.425)
Observations	16000	16000	16000	16000

Note: The average treatment effect is estimated using equation 4. Only the first treatment encountered for each participant is included. This table summarizes the across average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). 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: Phase 2 (Participant Fixed Effects)

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.723	0.630	0.372	44.551
	(0.003)	(0.004)	(0.004)	(0.385)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	0.027	-0.377	-0.221	-26.021
	(0.004)	(0.007)	(0.007)	(0.577)
No Disclosure	0.025	-0.429	-0.250	-29.759
	(0.004)	(0.007)	(0.007)	(0.604)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.035	0.064	0.046	3.749
	(0.005)	(0.006)	(0.007)	(0.593)
Stoplight	0.002	0.003	0.001	0.091
	(0.005)	(0.005)	(0.006)	(0.536)
Observations	80000	80000	80000	80000

Note: The average treatment effect is estimated using equation 4 with additional fixed effects at the participant level. This table summarizes the across average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). 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.7: Average Treatment Effects: Phase 2 (Controlling for Order)

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.727	0.677	0.430	53.221
	(0.005)	(0.009)	(0.010)	(0.840)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	0.027	-0.377	-0.221	-26.036
	(0.004)	(0.007)	(0.007)	(0.536)
No Disclosure	0.025	-0.429	-0.250	-29.739
	(0.004)	(0.007)	(0.007)	(0.576)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.034	0.063	0.045	3.690
	(0.005)	(0.006)	(0.006)	(0.532)
Stoplight	0.002	0.003	0.001	0.009
	(0.005)	(0.005)	(0.005)	(0.477)
Observations	80000	80000	80000	80000

Note: Note: The average treatment effect is estimated using equation 4 plus additional terms controlling for treatment order. This table summarizes the across average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). 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.8: Average Treatment Effects: Phase 2 (Controlling for Prior Statements Assessed)

Treatment	**Correct**	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)	(4)
Panel A: No Automation Baseline (β_0)				
Full Disclosure	0.728	0.681	0.429	53.345
	(0.004)	(0.004)	(0.004)	(0.340)
Panel B: Automation Treatment Effects (β_k)				
Full Disclosure	0.027	-0.377	-0.221	-26.028
	(0.004)	(0.004)	(0.004)	(0.383)
No Disclosure	0.025	-0.429	-0.250	-29.734
	(0.004)	(0.004)	(0.004)	(0.383)
Panel C: No Automation Treatment Effects (β_k)				
No Disclosure	-0.035	0.063	0.045	3.691
	(0.004)	(0.004)	(0.004)	(0.383)
Stoplight	0.002	0.003	0.001	0.022
	(0.004)	(0.004)	(0.004)	(0.383)
Observations	80000	80000	80000	80000

Note: Note: The average treatment effect is estimated using equation 4 plus additional terms controlling for the number of prior statements assessed. This table summarizes the across average treatment effects of different information environments on the assessment accuracy (column (1)), effort (columns (2)-(4)), and probability assessment (column (5)). 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.5.1 Squared Deviation from Ground Truth

Table A.9: Average Treatment Effects: Phase 2 (Deviation from Ground Truth)

Treatment	**Correct**	**Deviation from Ground Truth**
	(1)	(2)
Panel A: No Automation Baseline (β_0)		
Full Disclosure	0.723	0.338
	(0.004)	(0.003)
Panel B: Automation Treatment Effects (β_k)		
Full Disclosure	0.027	-0.005
	(0.004)	(0.003)
No Disclosure	0.025	0.001
	(0.004)	(0.003)
Panel C: No Automation Treatment Effects (β_k)		
No Disclosure	-0.035	0.032
	(0.005)	(0.003)
Stoplight	0.002	0.001
	(0.005)	(0.003)
Observations	80000	80000

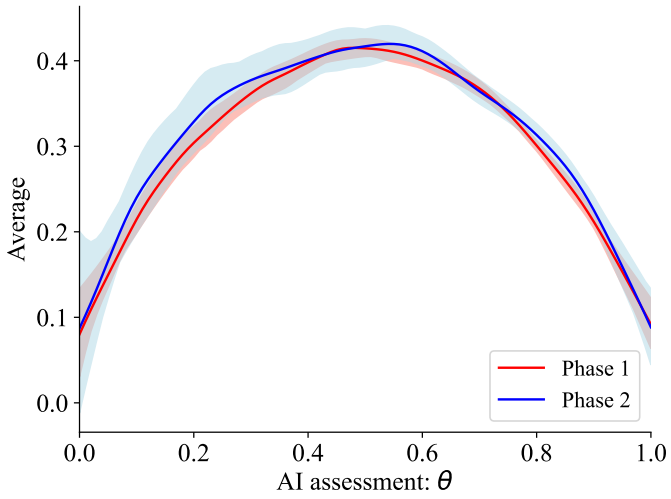
Note: The average treatment effect is estimated using equation 4. This table summarizes the across average treatment effects of different information environments on the assessment accuracy as measured by proportion correct (column (1)) and and deviation from ground truth (column (2)). In treatments 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: Phase 2 (Squared Deviation from Ground Truth)

Treatment	**Correct**	**Mean Squared Error**
	(1)	(2)
Panel A: No Automation Baseline (β_0)		
Full Disclosure	0.723	0.206
	(0.004)	(0.003)
Panel B: Automation Treatment Effects (β_k)		
Full Disclosure	0.027	-0.021
	(0.004)	(0.003)
No Disclosure	0.025	-0.022
	(0.004)	(0.003)
Panel C: No Automation Treatment Effects (β_k)		
No Disclosure	-0.035	0.021
	(0.005)	(0.003)
Stoplight	0.002	0.000
	(0.005)	(0.003)
Observations	80000	80000

Note: The average treatment effect is estimated using equation 4. This table summarizes the across average treatment effects of different information environments on the assessment accuracy as measured by proportion correct (column (1)) and 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.4: $V(\theta)$ Using Deviation from Ground Truth



Note: $V(\theta)$ is defined as $E\left[|p_{ij} - \omega_i| \ |\theta\right]$. $V(\theta)$ is estimated using local linear regression from the first phase 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.

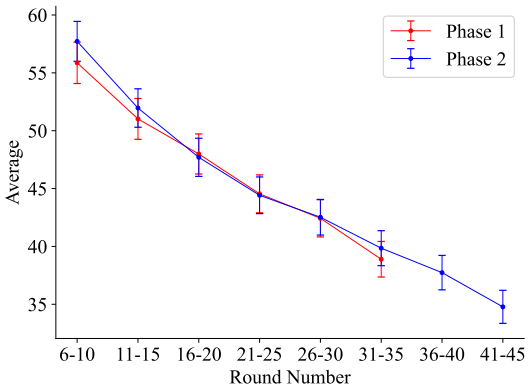
A.6 Fatigue and Learning

Participants classified 35 claims in phase 1 and 45 claims in phase 2. We test for fatigue and learning effects by estimating the following regression model and plotting β_k in figure A.5.

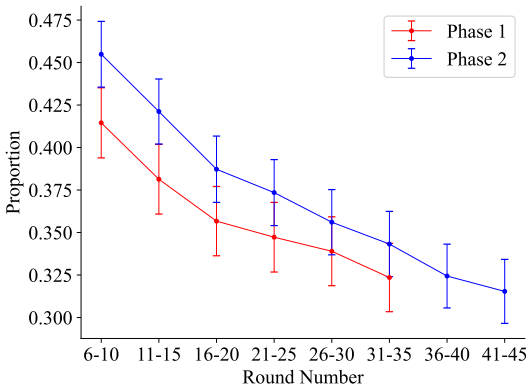
$$y_{i,j} = \sum_{k \in \text{ Intervals}} 1[\text{ interval } (i,j) = k]\beta_k + \sum_{k' \in \text{ Policies}} 1[\text{ policy } (i,j) = k']\gamma_{k'} + \varepsilon_{ij} \qquad (8)$$

Figure A.5: Outcome by Round Number

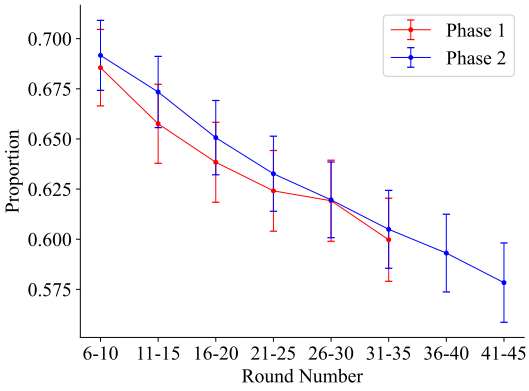
(a) Time



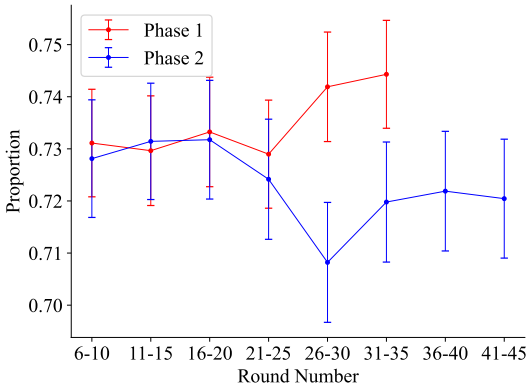
(b) Clicked Google Link



(c) Used External Sources



(d) Correct



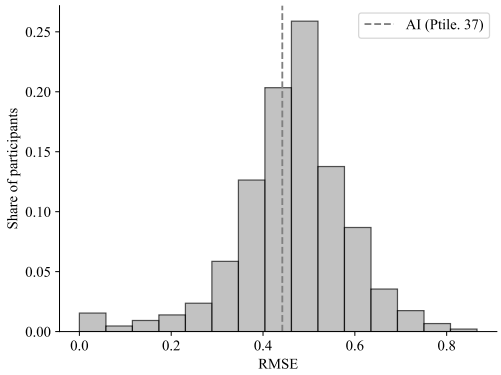
(e)

Note: Figure summarizes outcome by round number. For both phase 1 and phase 2, 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.

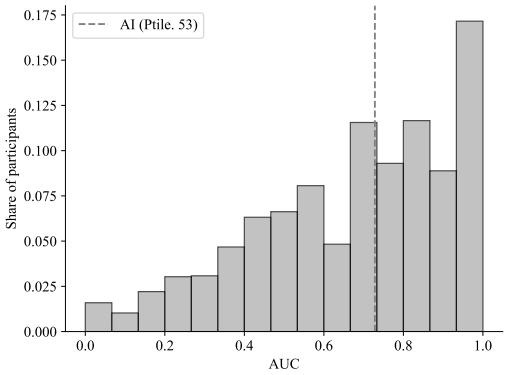
A.7 Comparing Accuracy

Figure A.6: Comparing AI Accuracy to Humans

(a) RMSE



(b) AUROC



Note: Figure A.6a compares AI accuracy to the distribution of participant accuracy in phase 2 under No Disclosure + No Automation as measured through mean-squared error. Figure A.6b compares AI accuracy to the distribution of participant accuracy in phase 2 under No Disclosure + No Automation as measured through AUROC.

A.7.1 Stability

Table A.11: Balance: Phase 1 vs Phase 2

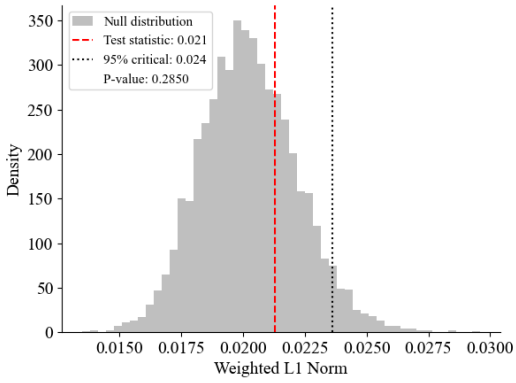
	Phase 1		**Phase 2**			
	Mean	SD	Mean	SD	Diff	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Correct Classification	0.735	0.441	0.723	0.447	0.012	0.008
Classified as True	0.696	0.460	0.696	0.460	-0.001	0.912
Assessment	0.630	0.329	0.629	0.318	0.001	0.732
Used External Sources	0.637	0.481	0.630	0.483	0.007	0.579
Clicked 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 phase one, while Columns (3) and (4) present the same statistics for phase two. 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 phase 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.

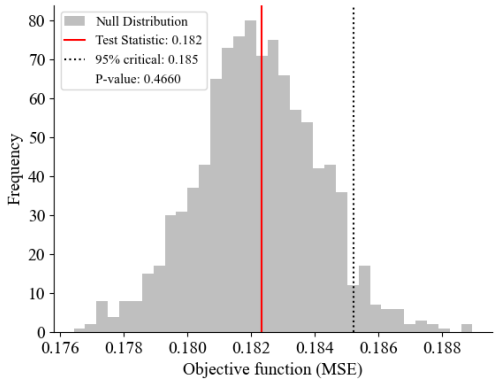
Figure A.7: Test for Stability and Convexity of $V(\theta)$ versus Null Distribution

a

(a) Stability

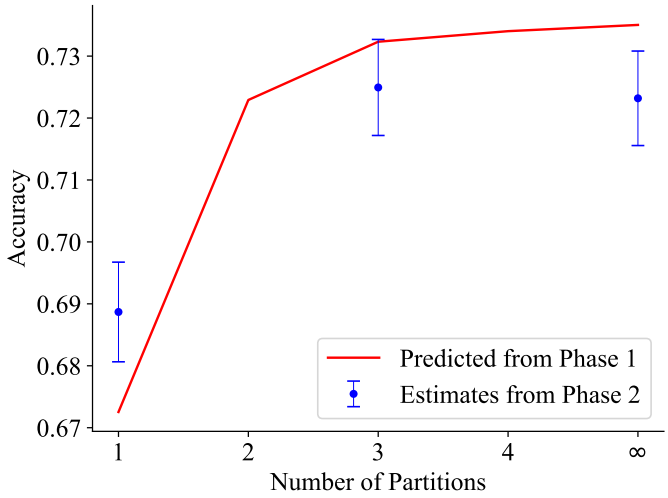


(b) Convexity



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 phase 1 and the $V(\theta)$ estimated from phase 2. Second, we construct a null distribution through bootstrap resampling: in each iteration, we randomly split the phase 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.

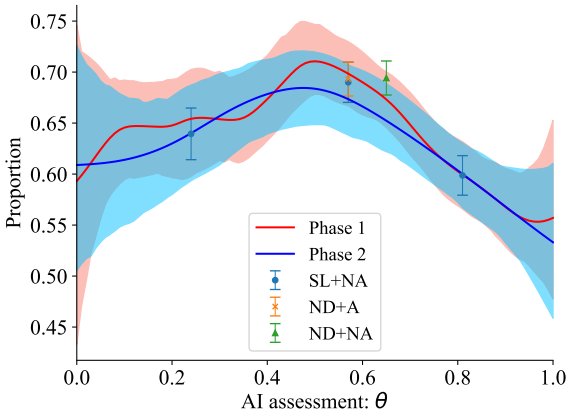
Figure A.8: Stoplight Policy Predicted Accuracy by K



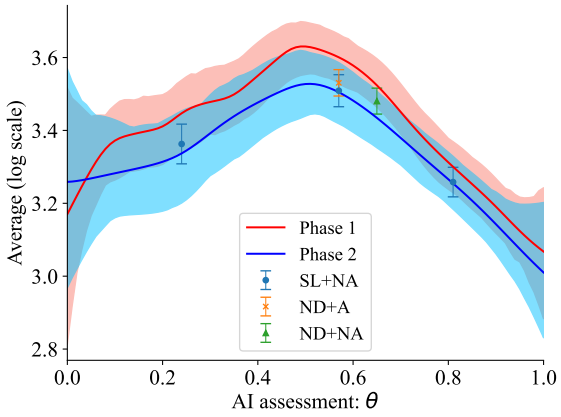
Note: Figure compares the predicted accuracy based on the model with the actual accuracy observed in the experiment. The estimated accuracy from phase 2 at $K = 1$ is the average accuracy in the No Disclosure + No Automation arm; $K = 3$ corresponds to the average accuracy in the Stoplight + No Automation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

Figure A.9: Stability of Effort Response

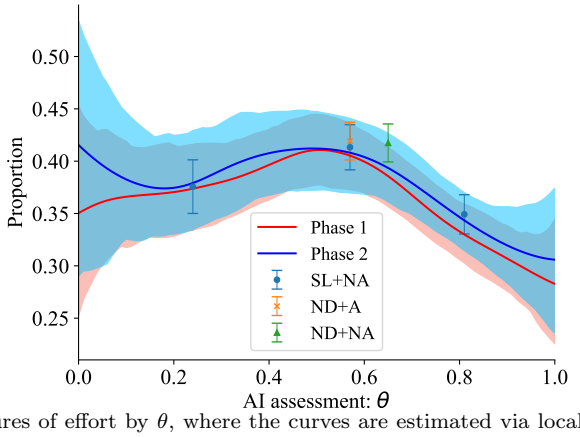
(a) Used External Sources



(b) Time Spent



(c) Clicked Google Link



Note: Figure summarizes measures of effort by θ, where the curves are estimated via local linear regression. For both phases, 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.

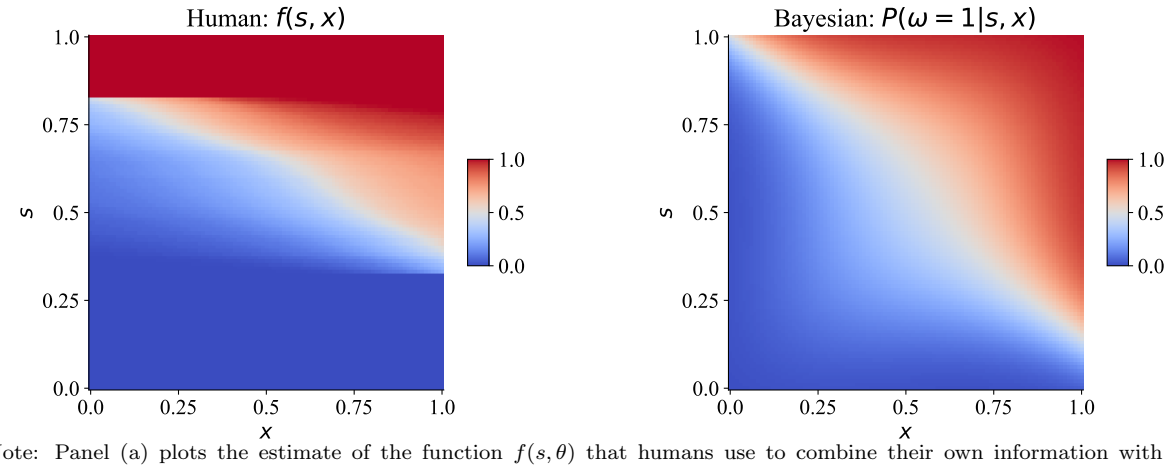
B Mechanism and Implications Appendix

B.1 Additional Empirical Results

Figure B.10: Human vs Bayesian Update Rule

(a) Humans

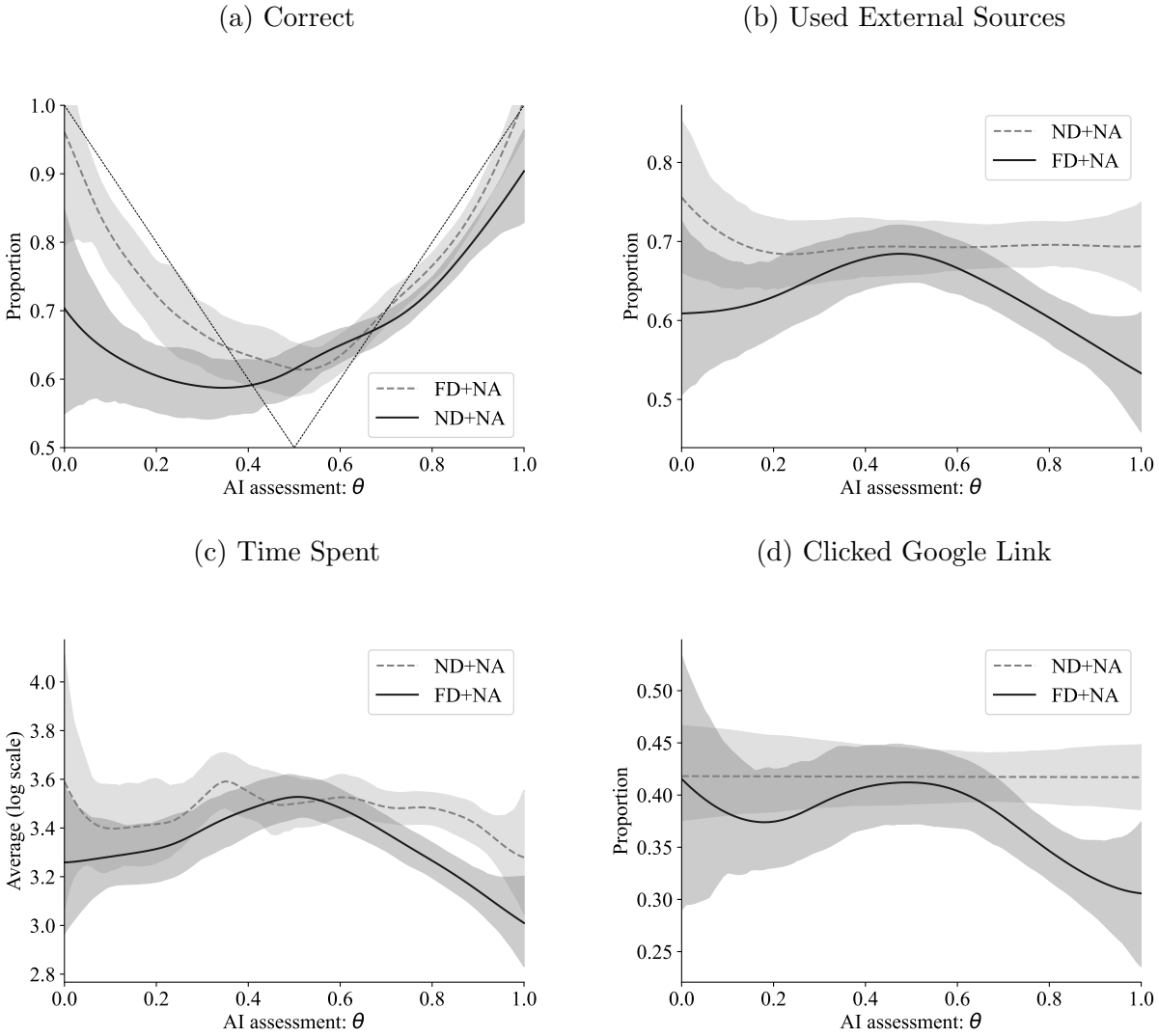
(b) Bayesian



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.

B.2 Heterogeneous Treatment Effects

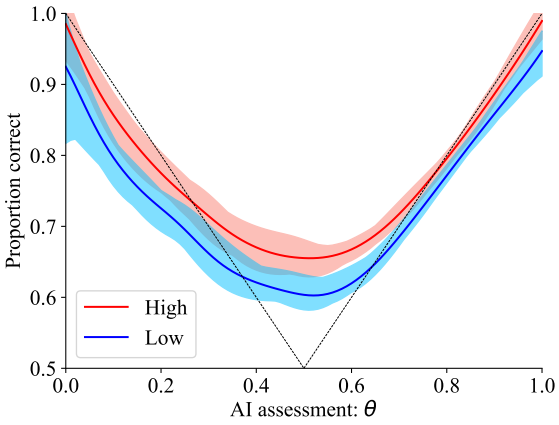
Figure B.11: Accuracy and Effort by θ



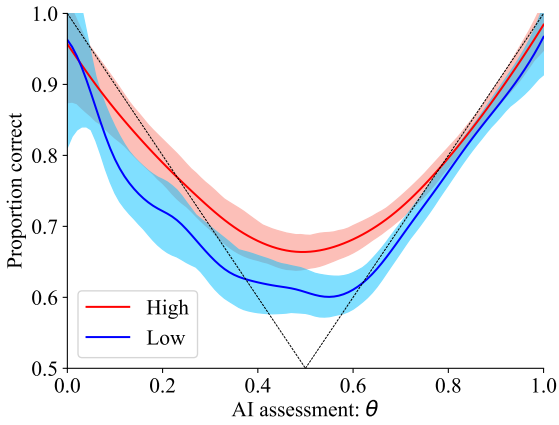
Note: Figure summarizes measures of effort by θ, where the curves are estimated via local linear regression. The figures only use data from phase 2. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Time Spent is the time spent 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 B.12: Heterogeneity in $V(\theta)$

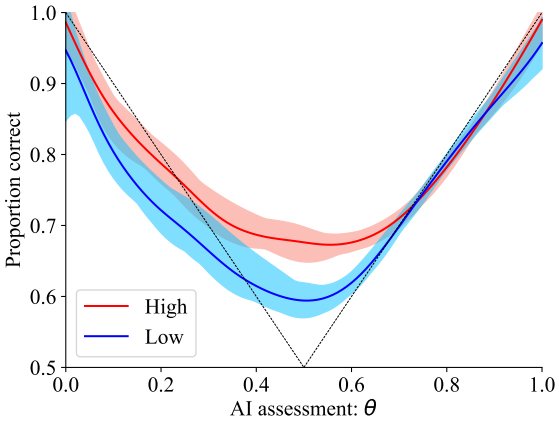
(a) Split by Comprehension



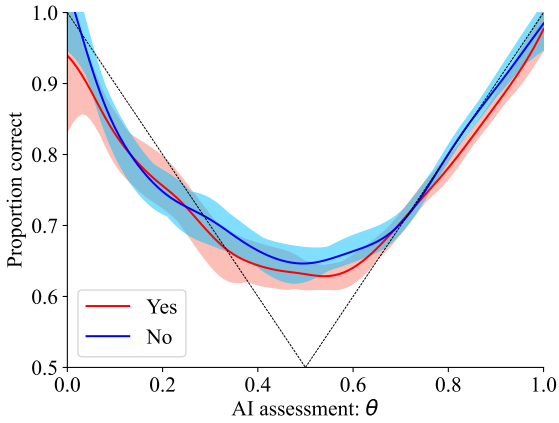
(b) Split by Performance



(c) Split by Effort



(d) Split by Overconfident



Note: $V(\theta)$ is estimated using local linear regression from the first phase data. $V(\theta)$ is estimated separately for high and low conscientiousness participants, and conscientiousness is measured in four ways: (1) number of comprehension questions answered correctly in the training section (two or less wrong indicates high conscientiousness), (2) performance as measured by a regression of correct minus $max\{\theta, 1-\theta\}$ on participant fixed effects, (3) effort as measured by a regression of used external sources indicator on θ, θ^2, and participant fixed effects, and (4) confidence as measured by a regression of the ground truth on a constant, and the probability reported interacted with participant fixed effects. For figures B.12b, B.12c, and B.12d, 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 B.12: Heterogeneity in Predicted Performance

	SL		FD+A		ND+A	
	Pooled	Separate	Pooled	Separate	Pooled	Separate
Comprehension						
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
Performance						
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
Effort						
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
Overconfident						
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 7) previously estimated on the standard $V(\theta)$ using all the phase 1 data. The separate column denotes the performance of individually estimated policies for each group (comprehension, performance, effort, and confidence) using the unique $V(\theta)$.

C Alternative Approaches to Design Human-AI Collaboration

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.

C.1 Algorithmic Triage Approach to Human-AI Collaboration

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 + No Automation) and requires additional data to predict performance.

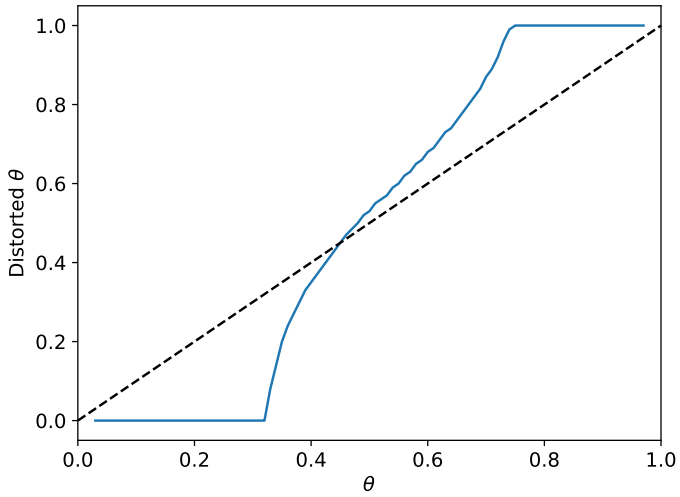
C.2 Exaggerating AI Signals to Overcome Automation Neglect

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 out 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 $\delta(\theta) : [0, 1] \to [0, 1]$ that maximizes

$$\theta T(\delta(\theta), 1) + (1 - \theta)(1 - T(\delta(\theta), 0)), \tag{9}$$

Figure C.13: Na ive 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.

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 $\omega = 1$, $\bar{\delta}(\tilde{\theta}) = \mathbb{E}[\theta|\delta(\theta) = \tilde{\theta}]$, leading to (sophisticated) expected accuracy $\mathbb{E}\left[V\left(\bar{\delta}(\theta)\right)\right]$.

It is straightforward to solve the na ive designer's problem and compare its na ive 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 C.13 plots the na ive 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 na ive optimal policy yields a na ive expected accuracy of 74.7%. This accuracy is very close to that under Full Disclosure + Automation (75.1%): intuitively, the na ive designer believes that she can nearly replicate automation by greatly 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 73.7% 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.

D Experimental Instructions

Below are the instructions the subjects received along with the interface-based treatment. These screenshots come from the second stage of the experiment. 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.

D.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

D.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

D.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

D.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

D.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

D.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