Designing Human-AI Collaboration: A Sufficient-Statistic Approach

Nikhil Agarwal, Alex Moehring, Alexander Wolitzky *

January 20, 2026

Abstract

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

JEL: C91, D83, D89, D47.

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

*First version: April 18, 2025. 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 Daron Acemoglu, David Atkin, Dean Eckles, Glenn Ellison, Guillaume Frechette, Drew Fudenberg, Joshua Gans, Anton Kolotilin, Benjamin Manning, Parag Pathak, Ashesh Rambachan, Frank Schilbach, Jesse Shapiro, Jann Speiss, and Tomasz Strzalecki, as well as seminar participants at MIT, Northwestern, and Purdue 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 pre-analysis plans are available at `www.socialscienceregistry.org/trials/13990`.

1 Introduction

Many predictive Artificial Intelligence (AI) tools now match or surpass human performance (Kleinberg et al., 2017; Agrawal et al., 2018; Lai et al., 2021). There is consequently great interest in how AI assistance affects human performance (Brynjolfsson et al., 2025) and in the design of human-AI collaborative systems that consider which cases to automate or to assign to humans, either with or without the assistance of AI predictions (Raghu et al., 2019; Mozannar and Sontag, 2020).

A challenge in designing human-AI collaboration is that the space of possible designs is large, and it can be difficult to predict how humans will respond to a design. Humans can exhibit biases in belief updating in response to AI predictions (Agarwal et al., 2023), and AI predictions 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). This complexity of possible responses, together with the dimensionality of the space of possible collaborative designs, frustrates the search for an optimal design via experimentation or structural modeling.

This paper develops a sufficient-statistic approach for designing human-AI collaboration for binary classification tasks, where each of several cases must receive a classification $a \in \{0, 1\}$.[1] The sufficient statistic, $V(x)$, is the probability that a human decision-maker correctly classifies a case when they observe a calibrated AI assessment that the probability that the ground truth is 1 is $x \in [0, 1]$.[2] We allow any AI system that selectively automates classification tasks based on its assessment and/or delegates tasks to a human decision-maker while disclosing a (potentially imperfect) signal of its assessment.[3] Under the assumption that V does not depend on the information disclosure policy, results from the literature on information design (Dworczak and Martini, 2019) imply that V can be used to find the optimal design in this space. That is, conditional on V, the optimal design does not depend on any other aspect of the human-AI interaction, such as humans' information, behavioral biases, or effort responses. Moreover, the function V can be readily estimated from data on decision accuracy when AI assessments x are fully disclosed to decision-makers. These data can be either experimentally generated (as in this paper) or previously observed.

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

[1] Our approach extends to multi-class classification problems and a range of designer objectives.

[2] This sufficient statistic is a function rather than a number as is typical in public finance (Chetty, 2009).

[3] We do not consider systems that first elicit the human's signal and then combine it with the AI assessment to reach a decision. Such a system would need to consider the possibility of strategic reporting or exaggeration by humans, making our approach a natural starting point. Moreover, we show in Section 5.2 that the accuracy gains from eliciting human signals are negligible relative to our optimal design when automation is feasible.

behavior and solves for the optimal design. This approach requires stronger behavioral assumptions. In addition, estimating such a model likely requires data on decision accuracy under varying AI assessments x, which suffices to directly estimate V. A second alternative experimentally tests several designs to find the optimal one. Such brute force search is impractical because the space of possible designs is large. Moreover, concluding that the best design tested is globally optimal would likely require stronger assumptions than ours.

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

Fact-checking is also conducive to experimental study. The task is easy to explain and can be conducted by untrained experimental participants. Measuring accuracy is straightforward, as there are established databases of true and false statements with curated ground-truth labels, such as FEVEROUS (Aly et al., 2021), which we use in our experiment. Fact-checking is also representative of other binary classification tasks, such as medical diagnosis (Agarwal et al., 2023), bail decisions (Kleinberg et al., 2017), and resume screening (Li et al., 2020).

Our experiment proceeds in two stages. The first stage estimates the sufficient statistic V by measuring classification accuracy on cases with different AI assessments and solves for optimal and approximately optimal designs. We consider both designs where automation is allowed and where humans make all classification decisions, as in many settings—potentially including fact-checking—there may be a societal preference for humans to make final decisions. In the second stage, we implement five designs derived from the first-stage estimates in a within-participant experiment. We test the sufficient-statistic approach by comparing the predicted classification accuracy from the first-stage to the second-stage results.

The first-stage results yields several insights. First, the estimated function V is convex. This property implies that fully disclosing the AI assessment is optimal for all cases that are delegated to human decision-makers. This finding contrasts with prior theoretical and empirical results (Athey et al., 2020; Dell'Acqua, 2022) that find that partially disclosing AI assessments can be optimal because disclosing more precise assessments crowds out human effort in information acquisition. While we also find effort crowding-out, this effect is too

weak to overturn the direct benefit of providing more precise AI information in our setting.

Second, when the disclosed AI assessment is confident (x is close to 0 or 1), humans' classification accuracy $V(x)$ is significantly lower than the accuracy under automation, which equals $\max\{x, 1-x\}$. This implies that humans under-respond to the AI assessment when updating their beliefs, because simply following the AI prediction would increase accuracy whenever $V(x) < \max\{x, 1-x\}$.[4] It also implies that automating these cases is optimal. Thus, the optimal design automates cases where the AI is confident and delegates the remaining cases to humans while providing them with the AI assessment. We call this policy **Full Disclosure + Automation** (**FDA**).

Third, because uncertain AI assessments add little value to humans' own assessments, we predict that a policy that automates cases where the AI is confident and delegates the rest to humans without AI assistance is approximately optimal. Thus, while both humans and AI add value (in particular, full automation is suboptimal), the value of direct human-AI *collaboration*—rather than selective automation and delegation—is negligible in our setting. Concretely, we predict that accuracy under FDA will be similar to that under **No Disclosure + Automation** (**NDA**), where we automate cases where the AI is confident, and delegate the rest to humans without AI assistance.

In addition, we predict that the optimal design when automation is infeasible is **Full Disclosure + No Automation** (**FDNA**), where humans are provided with the AI assessment. This design is predicted to significantly outperform **No Disclosure + No Automation** (**NDNA**), where humans do not receive AI assistance. Finally, we predict that accuracy under FDNA is very similar to that under a simpler **Stoplight** (**SL**) policy, where the AI communicates only one of three possible signals (e.g., "Likely False," "Uncertain," "Likely True," or "Red," "Yellow," "Green").[5]

The second stage experiment puts these predictions to the test. The predicted performance of all five policies is within 1.6 percentage points of the experimental estimates, and none of the differences are significant at the 1% level. Moreover, the qualitative predictions are all borne out: FDA is the best policy when automation is feasible but is statistically indistinguishable from NDA; and FDNA is the best policy when automation is infeasible but is indistinguishable from SL, while NDNA is significantly worse. This validation of our counterfactual predictions suggests that the sufficient statistic assumption is a good guide for designing human-AI collaboration in our context.

In addition to designing human-AI collaboration, we also explore the mechanisms that determine the shape of V (and hence the optimal designs and their accuracy). We decompose

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

[5] In the experiment, all signal realizations take the form of probability assessments, to avoid framing effects.

the impact of behavioral biases and effort crowd-out on classification accuracy. To provide a sharp lower bound on the impact of errors in belief updating, we compare the accuracy of AI-assisted humans with that of an optimal classifier based on both AI predictions and humans' reported assessments. We find that at least 7.7% of humans' incorrect classifications are attributable to errors in belief updating. Remarkably, the optimal FDA policy approximately achieves the optimal classifier benchmark, implying that there is little benefit to considering designs where humans' probability assessments can be communicated to the AI.

We next examine whether humans under-respond to AI information because they are overconfident in their own information or under-confident in the AI's. To do so, we estimate the update rule $p(s, x)$ that participants use to combine their private signal s with the disclosed AI prediction x to reach an assessment p.[6] We find that AI under-response is almost entirely due to overconfidence in own-signal precision: humans' beliefs are too sensitive to their own signals relative to a Bayesian benchmark but are appropriately sensitive to AI predictions. This result contrasts starkly with prior work that attributes AI under-response to under-confidence in AI signal precision (Agarwal et al., 2023).

Finally, we find that providing accurate AI information crowds out human effort, but the impact of this effect on the precision of humans' signals is small.

Related Literature

Comparing predictive AI tools and human decisions is an active area of research (Kleinberg et al., 2017; Mullainathan and Obermeyer, 2022). Several papers compare the accuracy of humans with AI assistance to either humans or AI alone (Angelova et al., 2023; Vaccaro et al., 2024). Rather than comparing humans and AI, our objective is optimally designing human-AI collaborative systems. This goal is shared with the "algorithmic triage" problem in computer science (e.g. Mozannar and Sontag (2020)) and with Raghu et al. (2019) and Agarwal et al. (2023) in economics. We highlight two key differences. First, these papers abstract away from endogenous changes in human beliefs or effort in response to the set of cases that are delegated or automated. However, both theoretical and empirical results suggest that effort crowding-out can be large when humans are assisted by AI tools (Athey et al., 2020; Dell'Acqua, 2022), and Appendix F.1 argues that endogenous belief responses are similarly important in our setting. Second, optimal collaboration design using these earlier approaches requires direct experimentation, because they lack a model for predicting

[6] We assume that our participants use a common update rule and that their signal distribution depends only on effort, the ground truth, and the AI assessment, and not directly on the disclosed AI assessment conditional on these variables. Section 6.2 shows that these assumptions allow us to identify $p(s, x)$ by combining data from the FDNA and NDNA treatments in the second stage.

accuracy under counterfactual AI assessments. In addition, none of these papers tests the performance of the optimal policy in a second-stage experiment.[7]

Our sufficient-statistic approach for predicting accuracy in counterfactual policies builds on insights from information design (Kamenica and Gentzkow, 2011). Our sufficient statistic $V(x)$ is the designer's indirect utility from inducing a posterior mean assessment x, as in Dworczak and Martini (2019). This "mean-measurable" design problem arises when the designer discloses information about a signal of a binary ground truth (Arieli et al., 2023), which we extend to the case where the decision-maker also observes a private signal. Like us, De Clippel and Zhang (2022) studies information design with a non-Bayesian receiver; however, they analyze the qualitative impact of particular belief distortions, while we estimate the designer's indirect utility $V(x)$ and apply standard information design arguments. Finally, a growing experimental literature tests the assumptions and predictions of information design (e.g., Fréchette et al. (2022)), rather than using it for optimal design. In addition to these differences, to our knowledge, this paper (along with Dreyfuss and Hoong, 2025) is the first to apply information design techniques to human-AI collaboration.

Dreyfuss and Hoong (2025) also takes an information design approach to designing human-AI collaboration. They restrict attention to designing monotone partitions for binary classification problems in a setting where the humans' information is a subset of the AI's. They find that a binary partition, which can be interpreted as a recommended binary action from the AI, is optimal and out-performs full disclosure. Since the humans' information is a subset of the AI's, automation would out-perform optimal collaboration if it were allowed. Instead, in our setting, humans add value: the optimal Full Disclosure + Automation design significantly outperforms full automation. In general, our framework predicts the performance of any automation and disclosure policy and thus identifies the globally optimal policy.

Some of our empirical results parallel findings in prior belief updating experiments. For instance, under-response to new information is a common finding in behavioral economics (Benjamin, 2019). We go further by showing that, in our setting, under-response to new information is driven by overconfidence in own-signal precision rather than under-confidence in the precision of the new information, using a novel definition and decomposition of over- or under-response to information (related to Augenblick et al. (2025)).

A notable feature of our study is the use of a two-stage experiment to construct a demanding test of the model, where the first stage estimates a sufficient statistic that is used to design an optimal policy, and the second stage validates the design. Other papers that conduct such an exercise include Misra and Nair (2011), Carrell et al. (2013), Dubé and Misra (2023), and

[7]McLaughlin and Spiess (2024) derive a minimax optimal AI recommendation algorithm in a potential-outcomes framework and test it in a one-shot experiment.

Ostrovsky and Schwarz (2023). Our approach is closest to Ostrovsky and Schwarz (2023), who derive a sufficient statistic using auction theory (the distribution of bidder valuations), estimate it in a first stage, solve for the optimal reserve price, and test it in a second stage. A qualitative difference from Ostrovsky and Schwarz (2023) is that the space of reserve prices is one-dimensional while the space of disclosure policies is infinite-dimensional, so our approach avoids an intractable task of experimenting over a large design space.[8] Not all prior tests of optimal design have been successful: Carrell et al. (2013) shows that assigning squadron peers in the US Air Force Academy to maximize performance failed because the policy design did not account for endogenous peer group formation. The possibility of such endogenous responses makes the type of counterfactual tests we conduct demanding.

2 A Framework for Human-AI Collaboration Design

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

2.1 A Sufficient Statistic

Each case i must receive a binary classification $a_i \in \{0, 1\}$ (e.g., False or True). The ground truth is denoted $\omega_i \in \{0, 1\}$, with prior $\Pr(\omega = 1) = \phi$. An AI tool produces an assessment $\theta_i \in [0, 1]$ of the probability that $\omega_i = 1$. The assessment is calibrated: $\Pr(\omega_i = 1|\theta_i) = \theta_i$.[9] 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 ground truth. For example, if the AI assessment is always perfectly accurate then $\theta_i \in \{0, 1\}$ with probability 1, while if the AI assessment contains no information then $\theta_i = \phi$ (the prior probability that $\omega_i = 1$)

[8] 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 infinite-dimensional. Misra and Nair (2011) estimates a structural model of dynamic effort allocation to design an optimal dynamic incentive contract and tests it in a second stage.

[9] The availability of a calibrated assessment is a fairly weak requirement, because a designer with paired data on uncalibrated AI assessments and the ground truth for a sample of cases can re-calibrate the AI. Similarly, the designer can re-calibrate the AI in response to (slow-moving) shifts in the data distribution.

with probability 1. In general, a better AI (one that provides more information about ω_i in the sense of Blackwell (1953)) corresponds to a more spread-out distribution F. We suppress the case subscript i for the rest of this 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 their own. Signals can potentially take any form, including quantitative statements like, "The AI assessment is $\theta = 0.7$," or qualitative ones like, "The AI assesses that the statement is likely true." Formally, the designer chooses an *automation/disclosure policy* $\sigma : \Theta \to \Delta(\{0,1\} \cup R)$, where R is an arbitrary set of signal realizations, and σ_θ is the probability that a case with AI assessment θ is either automatically classified as false ($\sigma_\theta(0)$), automatically classified as true ($\sigma_\theta(1)$), or delegated to a human-decision maker who receives signal r from the AI ($\sigma_\theta(r)$), for each possible $r \in R$. The designer's problem is to design an automation/disclosure policy σ to maximize the probability of correct classification, $\Pr(a = \omega)$.[10]

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

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

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

Assumption 1 can be viewed as combining two assumptions. First, there are no framing effects: signals matter only through their probabilistic content, not the language used to express them. This implies that the probability of correct classification when receiving a signal r depends only on the induced posterior probability distribution over AI assessments $\mu_r \in \Delta([0,1])$. Second, the posterior distribution over AI assessments μ_r affects the

[10] We consider alternative designer objectives in Appendix E.

probability of correct classification only through its mean $x = \mathbb{E}^{\mu_r}[\theta] = \Pr(\omega = 1|\theta \sim \mu_r)$. This implies that there is no benefit to disclosing a non-degenerate distribution over AI assessments rather than just the posterior x. For example, Assumption 1 requires that the probability of correct classification when the AI discloses that $\theta = 0.7$ must be the same as when the AI discloses that θ is a 50-50 mixture of 0.5 or 0.9.

Assumption 1 thus implies that a signal r can be identified with the induced posterior x, and a disclosure policy can be summarized as a distribution G of induced posteriors x. This property greatly simplifies the design problem, as well as our experimental design. In particular, a signal from the AI to our experiment's participants will take the form of a posterior $x = \mathbb{E}^{\mu_x}[\theta] = \Pr(\omega = 1|x)$. This form of communication is without loss under Assumption 1, and it simplifies both the description of the disclosure policy and the interpretation of the probabilistic content of signals for our experimental participants. In addition, this neutral presentation limits framing and experimenter demand effects in comparison to text signals.[11]

2.2 When Does Assumption 1 Hold?

In human-AI collaboration, humans combine AI signals with their own information to make decisions. We now explain when Assumption 1 holds in such a system. Here (and in our experiment), we assume that humans first receive a signal r from the AI, then acquire their own information s, and finally make a classification $a(s, r)$ on the basis of these two signals.

We first provide a reduced-form sufficient condition for Assumption 1, without positing a particular model of human belief updating or information acquisition.

Claim 1. *Assumption 1 holds if the human decision rule $a(s, r)$ and the distribution of the human signal $\Pr(s|r, \omega)$ both depend on the AI signal r only through the posterior x.*

Proof. According to hypothesis, $\Pr(a = \omega|r) = \sum_{(s,\omega)} 1\left[a(s, r) = \omega\right] \Pr(s|r, \omega) \Pr(\omega|r) = \sum_{(s,\omega)} 1\left[a(s, x) = \omega\right] \Pr(s|x, \omega)(x1[\omega = 1] + (1 - x)1[\omega = 0]) = V(x)$, where $x = \Pr(\omega = 1|r)$. □

A canonical example where this holds is when the decision-maker is a Bayesian with correctly specified beliefs who chooses $a(s, r)$ to maximize classification accuracy. In this case, if $\Pr(s|r, \omega)$ depends on r only through x (i.e., s is independent of r conditional on (x, ω)), then so does $\Pr(\omega|s, r)$, because

$$\frac{\Pr\left(\omega = 1|s, r\right)}{\Pr\left(\omega = 0|s, r\right)} = \frac{\Pr\left(\omega = 1|r\right)}{\Pr\left(\omega = 0|r\right)} \frac{\Pr\left(s|r, \omega = 1\right)}{\Pr\left(s|r, \omega = 0\right)} = \frac{x}{1 - x} \frac{\Pr\left(s|x, \omega = 1\right)}{\Pr\left(s|x, \omega = 0\right)}, \tag{1}$$

[11] The potential for leveraging "behavioral" framing effects and gains from unstructured human-AI communication (e.g., via an LLM) are important future directions (see Section 7).

and hence so does the optimal Bayesian decision.[12]

In addition, if the decision-maker chooses how much effort e to exert in acquiring information after observing the AI signal r, the chosen effort e and the signal s will both be independent of r conditional on (x, ω) if s is independent of r conditional on (x, ω, e). This gives a more structural sufficient condition for Assumption 1 for a Bayesian decision-maker.

Claim 2. *Assumption 1 holds if the decision-maker is a Bayesian with correctly specified beliefs who maximizes classification accuracy, and the distribution of the human signal conditional on effort* $\Pr(s|r, \omega, e)$ *depends on the AI signal* r *only through the posterior* x*.*

Proof. Under the hypothesis, $\Pr(a = \omega|r, e) = \sum_{(s,\omega)} 1\left[a(s, r) = \omega\right] \Pr(s|r, \omega, e) \Pr(\omega|r, e) = \sum_{(s,\omega)} 1\left[a(s, x) = \omega\right] \Pr(s|x, \omega, e)(x1[\omega = 1] + (1 - x)1[\omega = 0])$, which depends only on (x, e). This implies that the optimal e depends on r only through x, and hence, by (1), so does $\Pr(a = \omega|r)$. □

Assumption 1 also holds if decision-makers make errors in probabilistic reasoning, but nonetheless their response to AI signals depends only on the posterior x. For example, it holds for a decision-maker who uses a non-Bayesian procedure such as weighted linear or non-linear averaging. A leading example is the belief-updating model in Grether (1980): a Grether agent updates according to (1) with heterogeneous exponential weights on the ratios $x/(1 - x)$ and $\Pr(s|x, \omega = 1)/\Pr(s|x, \omega = 0)$, so the resulting posterior belief again depends on r only through x.

In contrast, Assumption 1 is typically violated with conditionally dependent private signals, as then $\Pr(s|r, \omega)$ is unlikely to depend on r only through x. For example, if the human signal s and the AI assessment θ are perfectly correlated, then human decision accuracy following a signal r that reveals that $\theta = 0.5$ is 0.5 (as then s also equals 0.5), while human decision accuracy following a signal r that reveals that θ is a 50-50 mixture of 0 and 1 is 1 (as now s is either 0 or 1), even though these two signals give the same posterior $x = 0.5$.

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

Our approach remains valid if the designer's indirect utility V differs from the probability of correct classification, so long as Assumption 1 holds with the relevant V. For example,

[12] In general, for a correctly-specified Bayesian decision-maker whose decision rule depends on (s, r) only through $\Pr(\omega = 1|s, r)$, Assumption 1 holds if and only if s depends on x linearly conditional on ω, so that $\Pr(s|x, \omega) = (1 - x) \Pr(s|0, \omega) + x \Pr(s|1, \omega)$ for all (s, x, ω).

if the designer's objective is minimizing the expected deviation of a human decision-maker's probability assessment from the ground truth, then a would be a probability assessment in $[0,1]$ and $V(x)$ would be (minus) the expected deviation conditional on the AI disclosing that $\Pr(\omega = 1) = x$. For another example (inspired by Dreyfuss and Hoong (2025)), if the designer's objective is motivating humans to follow the AI's signal, $V(x)$ would be the probability the human takes $a = 0$ if $x < 0.5$ or $a = 1$ if $x > 0.5$. We calculate the optimal design under these alternative objectives in Appendix E.[13] The designer's objective can also weigh other outcomes (such as effort), so long as they are observable and satisfy Assumption 1.[14]

Finally, while we focus on binary classification problems, a similar approach applies for multi-class problems. In the general multi-class case with n possible classifications, the ground truth ω lies in an arbitrary finite set Ω with n elements. The generalization of Assumption 1 is that the probability of correct classification when receiving a signal r depends only on the posterior distribution over ω, $\Pr(\cdot|r) \in \Delta(\Omega)$. Under this assumption, the designer's problem becomes a multi-dimensional *moment persuasion* problem, as formulated in Dworczak and Kolotilin (2024).[15] The main difference from the binary case is that the indirect utility function to be estimated and the set of possible disclosure policies to be optimized over are both lower-dimensional in the binary case.

2.3 The Designer's Problem

We now explain how to find the optimal automation/disclosure policy from estimates of the indirect utility function V. Under Assumption 1, an information disclosure policy can be summarized by the distribution G of induced posteriors x. A key result from the information design literature (Blackwell, 1953; Gentzkow and Kamenica, 2016; Kolotilin, 2018) 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 F of AI assessments θ. Therefore, the maximum expected accuracy attainable by information disclosure alone (without automation) is

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

[13] The optimal design in the latter case is a binary signal, as in Dreyfuss and Hoong (2025), and full disclosure in the former case.

[14] Assumption 1 also does not require that the decision-makers' objective is maximizing classification accuracy. It requires only that the designer's indirect utility $V(x)$ is a well-defined function of the posterior, not that $V(x)$ is also the decision-maker's utility.

[15] Here, the designer's indirect utility is a function $V : \Delta(\Omega) \to \mathbb{R}$ defined on the $n-1$-dimensional simplex, and the designer maximizes $\int_{\mu \in \Delta(\Omega)} V(\mu) d\tau(\mu)$ over disclosure policies $\tau \in \Delta(\Delta(\Omega))$, subject to the Bayes plausibility constraint $\tau \in MPC(\phi)$, where $\phi \in \Delta(\Delta(\Omega))$ is the distribution of the AI assessment.

where $MPC(F)$ denotes the set of all distributions that are mean-preserving contractions of the distribution of AI assessments F. For example, under the *full disclosure* policy, where the AI always discloses its assessment, expected accuracy equals $\int_0^1 V(x)\,dF(x)$; while under the *no disclosure* policy, where the AI reveals no information, expected accuracy equals $V\left(\int_0^1 x dF(x)\right) = V(\phi)$.

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

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

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

If the human decision-maker is a correctly-specified Bayesian, then $V(x) \geq \max\{1-x, x\}$, because $\max\{1-x, x\}$ is the accuracy of a Bayesian 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 humans are irrational or under-respond to information provided by the AI, 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.

To calculate the optimal policy in practice, the designer must estimate the distribution of calibrated AI assessments F and the function $V(x)$ describing human decision accuracy as a function of the disclosed posterior x. In our experiment, the distribution of assessments F is given and known. Our experiment collects data to estimate the function $V(x)$, although observations from fully disclosed AI assessments would suffice in other context. We then calculate the optimal automation/disclosure policy and the optimal disclosure-only policy as described above. One can also calculate policies that are personalized to observed heterogeneity across decision-makers or cases by estimating V conditional on these observables.

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 selective automation but no disclosure of AI assessments on cases that are delegated to humans. In this problem, the designer chooses a set of AI assessments $\Theta^{\text{aut}} \subset [0,1]$ to automate, and the remaining cases with assessments $\theta \notin \Theta^{\text{aut}}$ are delegated to a human, who is informed only of the posterior

among delegated cases, $\mathbb{E}\left[\theta \,|\theta \notin \Theta^{\text{aut}}\right]$. The set Θ^{aut} solves

$$\max_{\Theta^{\text{aut}} \subset [0,1]} \mathbb{E}\left[\max\{\theta, 1-\theta\} \,\middle|\, \theta \in \Theta^{\text{aut}}\right] \Pr(\theta \in \Theta^{\text{aut}}) + V\left(\mathbb{E}\left[\theta \,\middle|\, \theta \notin \Theta^{\text{aut}}\right]\right) \Pr(\theta \notin \Theta^{\text{aut}}).^{16} \quad (4)$$

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

2.4 Discussion of the Optimal Design

We now describe how the shape of the function V determines the optimal automation/disclosure policy. First, full disclosure without automation is optimal if and only if V is convex and $V(x) \geq \max\{1-x, x\}$ for all x. These conditions hold, for example, if the human decision-makers are Bayesian with fixed effort and conditionally independent private signals.[17]

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

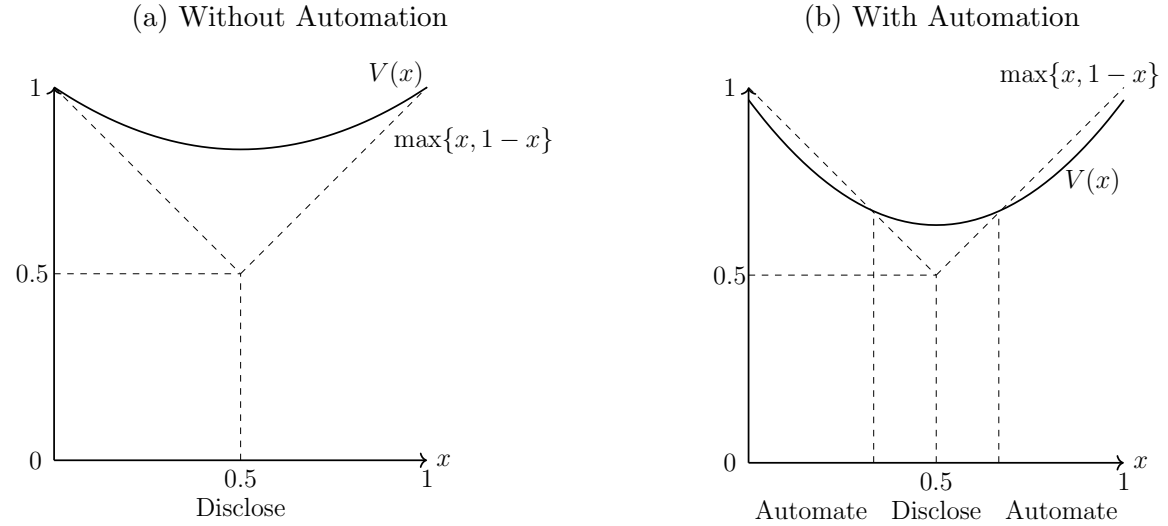
To preview, we will estimate that in our setting V is (approximately) convex, so fully disclosing the AI assessment is optimal. We also find values of x where $V(x) < \max\{1-x, x\}$, so automation is valuable. Qualitatively, our estimated function V has a similar shape as the function V in Figure 1b.

In other settings, different disclosure policies can be optimal. In particular, the function

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

[17] 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 their 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 is Optimal



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), delegation with 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 $V(x) < \max\{1-x, x\}$.

V may be non-convex if human effort is sufficiently sensitive to x. Figures 2a and 2b illustrate such functions. 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. Dell'Acqua (2022) finds an empirical setting where human effort is sufficiently sensitive to the disclosed AI signal that overall accuracy is higher with a less precise AI signal, which would imply that V is non-convex under Assumption 1. Agarwal et al. (2023) likewise finds that withholding the AI signal improves accuracy for some cases.

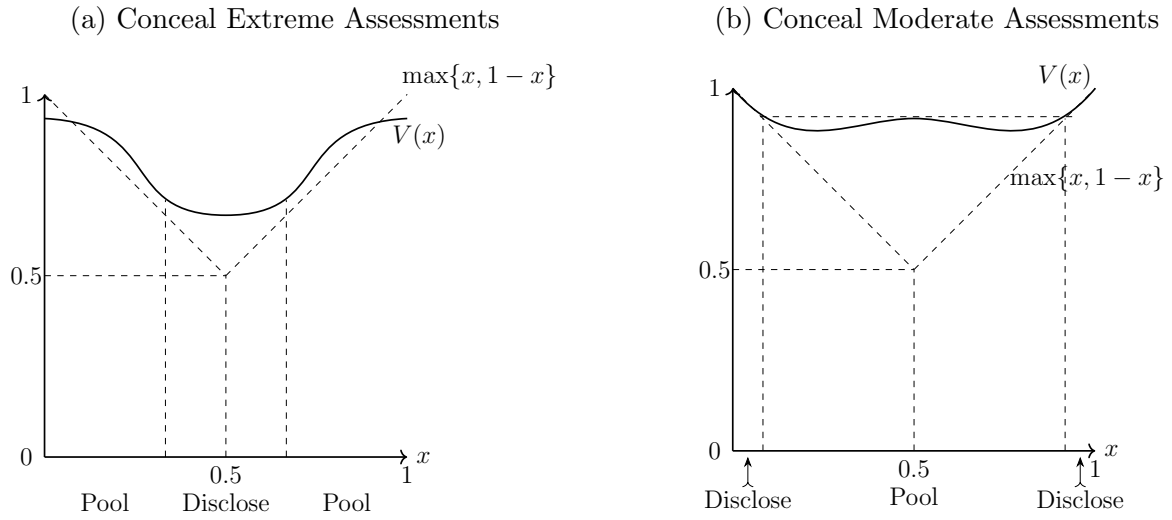
2.5 What if Assumption 1 is Violated?

Assumption 1 is unlikely to hold exactly in practice — we find evidence for most predictions of Assumption 1, but also a few violations (see Section 5.1). If Assumption 1 is violated, the following is a weaker assumption that still allows an information design approach.

Assumption 1'. *The probability that a human decision-maker correctly classifies a case when they receive a signal r from the AI depends only on the induced posterior probability distribution over AI assessments, $\mu \in \Delta([0,1])$. We denote the probability of correct classification at distribution μ by $\tilde{V}(\mu)$.*

Assumption 1' retains from Assumption 1 the requirement that signals matter only through their probabilistic content, but drops the requirement that the posterior distribution over AI assessments matters only through its mean.

Figure 2: Indirect Utilities where Partial Disclosure is Optimal



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

In general, optimal design under Assumption 1' is much more difficult than under Assumption 1, because the relevant sufficient statistic is now the infinite-dimensional function $\tilde{V}(\mu)$, rather than the one-dimensional function $V(x)$.[18] However, if $\tilde{V}$ is convex, then the optimal design is the same as that under Assumption 1 with a convex V: disclose assessments θ where $\tilde{V}(\delta_\theta) \geq \max\{1-\theta, \theta\}$, and automate the decision if $\tilde{V}(\delta_\theta) < \max\{1-\theta, \theta\}$. In Section 5.1, we will describe how, for this reason, the data indicate that the optimal policy we find under Assumption 1 is robust to relaxing Assumption 1 to Assumption 1'.

3 Experimental Design

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

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

[18] That is, the information design problem is now the general one of Kamenica and Gentzkow (2011), rather than the mean-measurable problem of Dworczak and Martini (2019).

automation), No Disclosure + No Automation, and Stoplight.

We pre-registered this two-stage design and updated the plan to describe the specific policies tested in Stage 2 as a result of the Stage 1 estimates.[19] The experiment was implemented on Prolific (`www.prolific.com`) using an interface designed on the o-tree framework (Chen et al., 2016) that can be accessed through a browser.

3.1 The Task

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

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

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

In Stage 1, each participant assesses 30 randomly drawn statements from our database (described in Section 3.3). In Stage 2, each participant assesses 40 randomly drawn statements: eight different statements under each of the five policies. We use a within-participant design to maximize statistical power. We randomize the order of the policies to ensure that our estimated treatment effects are not confounded with learning or fatigue and to preserve a robustness check using a pure across-participant comparison based on the first treatment.

3.2 Participant Recruitment, Training, and Incentives

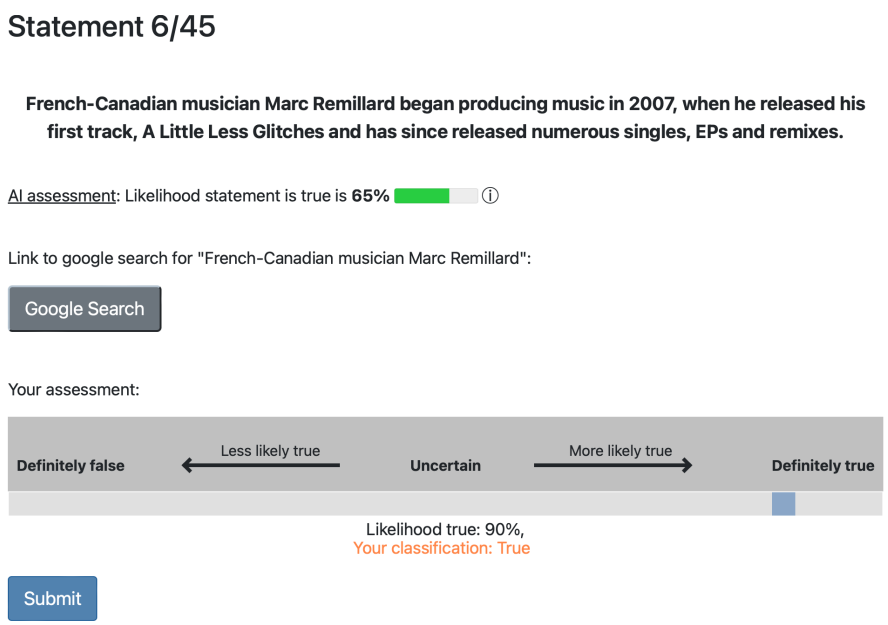
We recruit participants from Prolific, ensuring non-overlapping participant pools in each stage. We recruit a sample representative of the United States adult population on sex, age, and ethnicity.[20] Appendix Table A.2 summarizes the participants' demographic informa-

[19] We also pre-registered that we would update the plan after Stage 1 with the Stage 2 policies we test. The updated pre-registration changed the structure of the second stage to test 5 policies rather than the 4 we initially intended to test. We also reduced the number of statements per policy to 8 rather than 10 to maintain the overall duration of the experiment for each participant. Unless otherwise noted, all analyses we present are pre-registered.

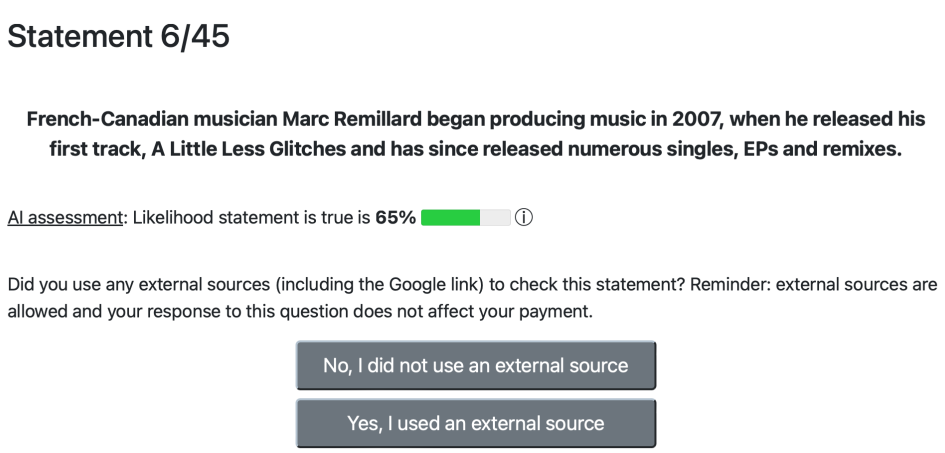
[20] Some segments are under-represented on Prolific, particularly older adults. We maintained the representative target until 95% of slots were filled and filled the remaining slots with non-representative participants.

Figure 3: Screenshots of Experimental Interface

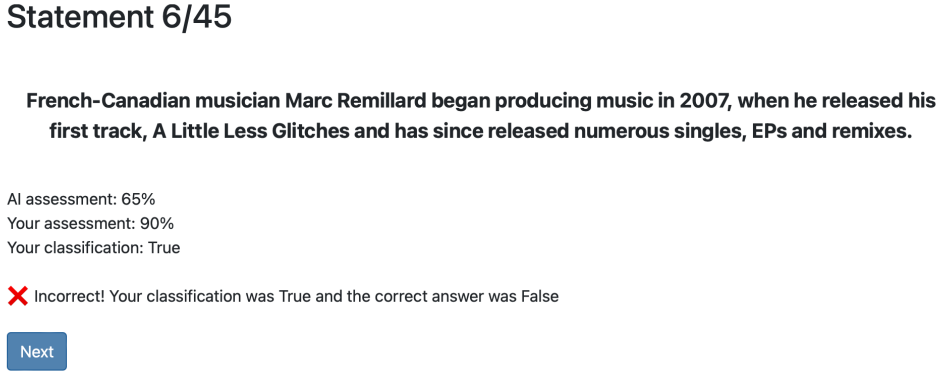
(a) Assessment Screen



(b) Self-Reported Effort Screen



(c) Feedback Screen



tion. Attrition was minimal, with 97.7% of participants who consented and began the study completing Stage 1 and 95.8% completing Stage 2.

At the start of the study, participants receive an overview of the task and the compensation rule. They then receive additional information about the task, the interface, and the share of true statements in the database. We next introduce the AI fact checker, explaining that it provides a calibrated assessment of the likelihood that a statement is true. In Stage 2, we explain that participants will encounter multiple AI fact-checkers (see Appendix I.4).

Next, we explain the compensation rule in broad terms and highlight that the expected payment increases with the accuracy of the assessment. Participants are incentivized in two ways to exert effort and provide accurate assessments. The first is a bonus of 35 cents for each correctly classified statement. The second is a lottery for an additional $20, where the win probability increases with the accuracy of the participant's assessments following Hossain and Okui (2013). The combined incentive scheme is a proper scoring rule. The detailed rule is available to participants at the click of a button.

The participants then encounter comprehension questions that check whether they understand that the AI is calibrated, that they can use outside resources, and the compensation rule. Each participant also assesses five incentivized practice statements to ensure familiarity with the experimental interface and these were excluded from all analyses. The full experimental instructions are presented in Appendix I.

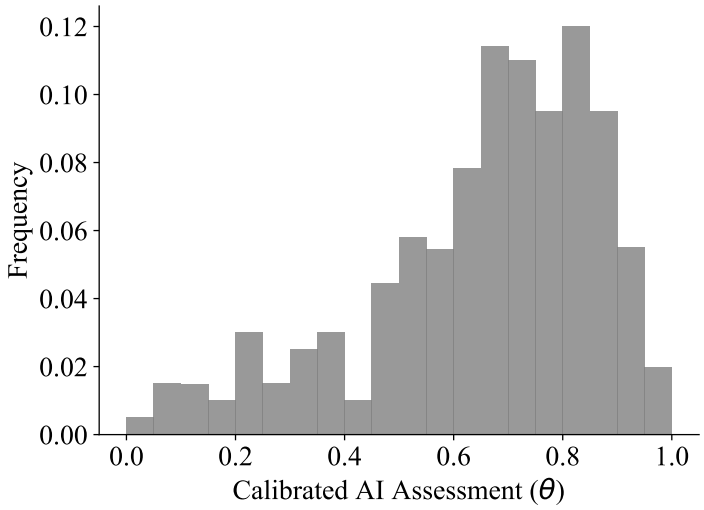
3.3 The Statements

We use the statements collected and labeled in the FEVEROUS database (Aly et al., 2021). The database contains approximately 80,000 statements that are generated by annotators from snippets of highlighted Wikipedia text or tables. A separate set of annotators then labels each statement as either Supported (True), Refuted (False), or Not Enough Information (NEI).[21] The statements and labels underwent extensive quality controls (Aly et al., 2021). In addition, we remove statements that are not suitable for our study. We remove the approximately 3% of statements with an NEI label; statements with spelling or grammatical errors flagged by either the rules-based LanguageTool API or GPT-4o;[22] and statements that we determine to be of poor quality, which are mostly statements where the ground truth changes over time, such as statements that reference an individual's age. In the final

[21]Supported statements require all claims in the statement to be verified and supported by evidence. A statement is refuted if any claim it contains is contradicted by evidence. Statements without sufficient evidence on Wikipedia for an unambiguous determination are labeled Not Enough Information.

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

Figure 4: Distribution of Calibrated AI Assessments



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

database of 41,969 statements, 65.4% of statements are True.[23]

3.4 The AI Fact-Checker

We generate calibrated AI assessments using OpenAI's GPT-4o without access to the internet. GPT-4o generated more accurate assessments than other alternatives, including the fact-checker in Aly et al. (2021). For each statement, we query the OpenAI API with the prompt, "True or False: [statement]" and retrieve the top 20 most likely next tokens along with the probability of each token. We calculate a raw score θ_i^r for each statement i as

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

where token_{ij} is the j^{th} most likely next token and p_{ij} is the probability assigned to it.[24] We then calibrate θ_i^r by binning it into 200 bins and calculating the share of true statements in each bin to yield the calibrated AI assessment θ_i. The calibrated assessment θ_i is approximately monotone in the raw score θ_i^r. Figure 4 shows the distribution of θ_i.

The AI assessments θ may differ from our participants' assessments because humans and

[23]Our review of 50 randomly drawn statements, half of which are true, found three cases in which our assessed label differed from the label in FEVEROUS and three cases where there was not enough information to assess the ground truth.

[24]GPT-4o is highly likely to suggest tokens in the set $\{\text{true}, \text{false}\}$. In our sample, the probability that the next token is not "true" or "false" $(1 - \sum_j p_j 1\left[\text{token}_j \in \{\text{true}, \text{false}\}\right])$ is less than 1% for 94.7% of statements.

GPT-4o may interpret statements differently or may have access to different information, particularly because the GPT-4o API we used did not have access to the internet (although the model may have implicitly memorized some of the relevant evidence). Similar considerations are relevant for prospective fact-checking and other classification tasks.

4 Stage 1 Results

Stage 1 estimates the function V and calculates optimal and approximately optimal information disclosure policies with and without automation. We also calculate the predicted treatment effect of each policy and document how effort responds to the AI assessment.

4.1 Overall Accuracy and Effort

Table 1 describes participants' accuracy and effort. Participants correctly classified 73.5% of statements. This overall accuracy is similar to the accuracy of 73.3% that would result if participants parroted the AI assessment, which is fully disclosed in Stage 1. However, this similarity masks large differences in accuracy $V(x)$ across AI assessments x. This heterogeneity is key for determining the optimal automation/disclosure policy and is discussed in the next subsection.

Participants classified 69.6% of cases as True and the average assessment was 63.0%, which is close to the share of true cases. Participants reported using external information sources in 63.7% of cases; clicked the provided Google search link in 36.0% of cases; and took an average of 46.8 seconds to fact-check each statement.[25]

4.2 Accuracy and Effort by AI Assessment x

Figure 5 presents our estimate of the sufficient statistic V, obtained using a local linear regression. The estimated function $\hat{V}$ has a qualitatively similar shape as the function V in Figure 1b. There are two important features.

First, $\hat{V}$ is approximately convex, and a statistical test does not reject that V is convex (p =0.96).[26,27] Recall that if V is convex then fully disclosing the AI assessment is optimal for all non-automated cases. We thus obtain a key implication for optimal design: in any

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

[26]We use the bootstrap procedure of Fang and Seo (2021) to test convexity of V (Figure B.3).

[27]Figure E.8 contains an estimate of V when the designer's objective is to minimize the deviation of the probability assessment from the ground truth ($V(x) = -E\left[|p_{ij} - \omega_i| \mid x\right]$). We also find V to be convex for this alternative objective.

Table 1: Stage 1 Summary Statistics

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

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

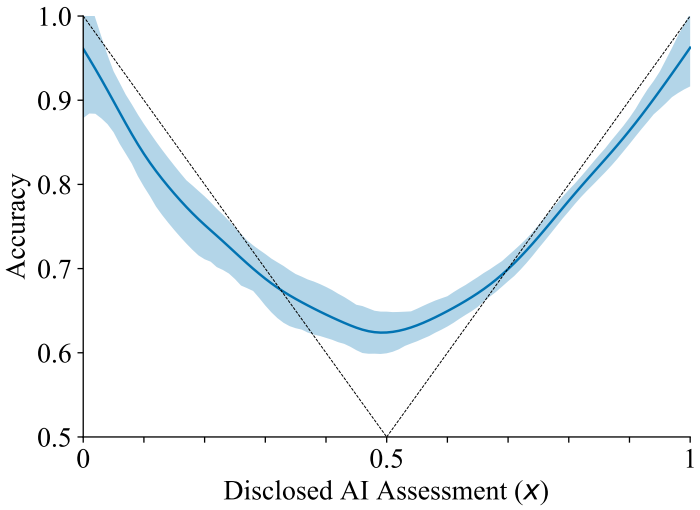
optimal automation/disclosure policy, the AI assessment of any non-automated case should be fully disclosed to the human decision-maker.

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

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

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

Figure 5: First Stage Estimate of V



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

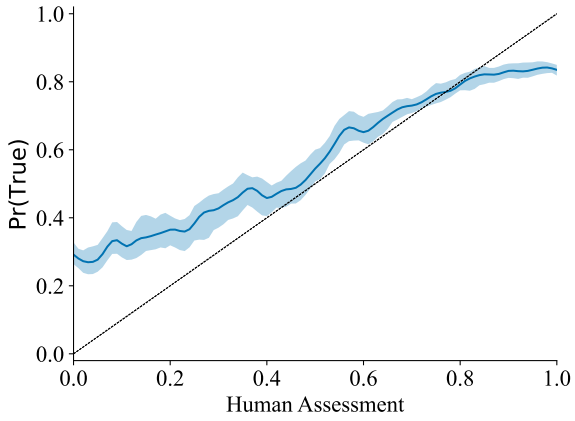
with correct beliefs about the precision of the AI signal but erroneously high beliefs about the precision of their own signal.

Examining the participants' reported assessments suggests overconfidence. Figure 6a plots the calibration curve (true probability against reported probability) for Stage 1 participants. The slope of the calibration curve is less than 1, indicating overconfidence. For example, 29% of statements that participants report are definitely false (reported $p = 0$) are actually True, and 16% of statements that participants report are definitely True (reported $p = 1$) are actually False. This miscalibration is particularly stark among individuals who do not exert effort as measured through self-reported use of external sources. While Figure 6a suggests overconfidence, it does not quantify its impact or speak to automation neglect. We address these questions using a structural model of belief updating in Section 6.

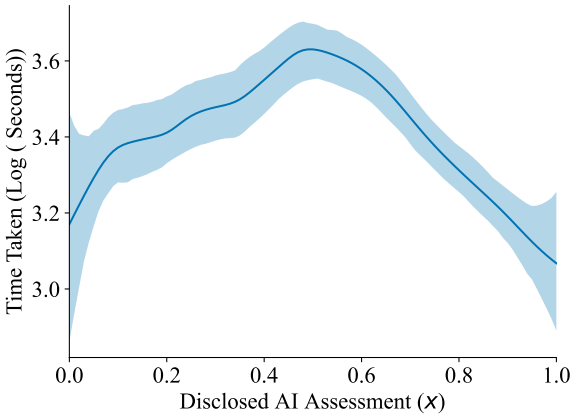
We also find evidence of effort crowding-out as the AI assessment x moves away from 0.5, the point of maximum uncertainty. Figure 6b shows that time taken is approximately 40% lower when $x = 1$ as compared to $x = 0.5$. This effect is similar for our other effort measures (see Figures A.1a and A.1b). This reduction in effort for confident AI assessments reduces performance under human-AI collaboration. In Section 6.4, we estimate the effect of disclosing the AI assessment on the precision of participants' private signals via the induced reduction in participant information-acquisition effort.

Figure 6: Miscalibration and Effort Response

(a) Calibration Curve of Human Assessments



(b) Time Taken by AI Assessment



Note: Figure 6a shows the 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. Figure 6b shows the log time taken (in seconds) to assess a statement by x in Stage 1, estimated via local linear regression. The 95% uniform confidence bands are computed via bootstrap accounting for clustering at the participant and case level.

4.3 Optimal and Simple Policies

We now solve for the optimal policy, both when automation is feasible and when it is not. That is, we solve problems (2) and (3) for the estimated function V. Since the estimated function V is convex, optimal policies fully disclose the AI assessment of any non-automated case. We compare the optimal policies with the optimal no-collaboration policies where the AI discloses no information to the human decision-maker: that is, the No Disclosure + No Automation policy and the No Disclosure + Automation policy that solves problem (4). We also consider the Stoplight policy where the AI can only disclose one of three signals. In total, we consider the four policies illustrated in Figure 7 in addition to Full Disclosure + No Automation (the optimal policy without automation), which was also run in Stage 1.

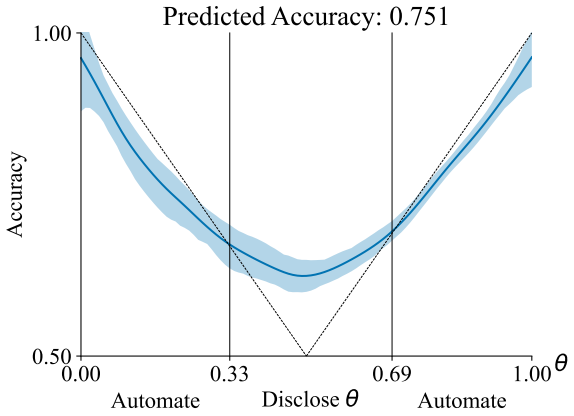
The first two policies allow automation. Here we compare the optimal policy (Full Disclosure + Automation) and the optimal no-collaboration policy (No Disclosure + Automation).

(a) **Full Disclosure + Automation** (**FDA**): The optimal policy (i.e., the solution to problem (3)) discloses θ if $V(\theta) > \max\{\theta, 1-\theta\}$—which we find holds if $\theta \in [0.33, 0.69]$—and automates the case otherwise. The predicted accuracy of this policy is 75.1%.

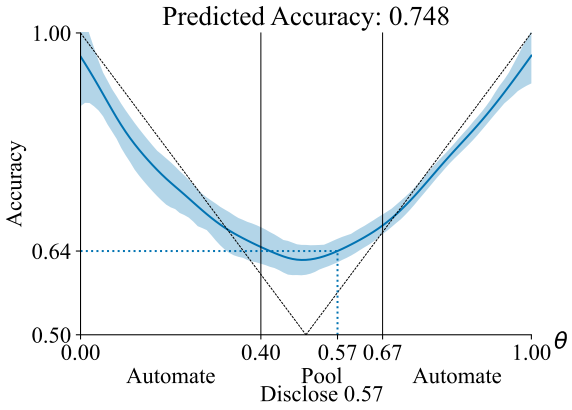
(b) **No Disclosure + Automation** (**NDA**): The optimal no-collaboration policy (i.e., the solution to problem (4)) automates cases in the set $\Theta^{\text{aut}} = [0, 0.39] \cup [0.68, 1]$ and otherwise discloses the share of true cases conditional on $\theta \notin \Theta^{\text{aut}}$, which equals 0.57. The predicted accuracy of this policy is 74.8%. Since this is only 0.3 percentage

Figure 7: Stage 2 Experiment Overview

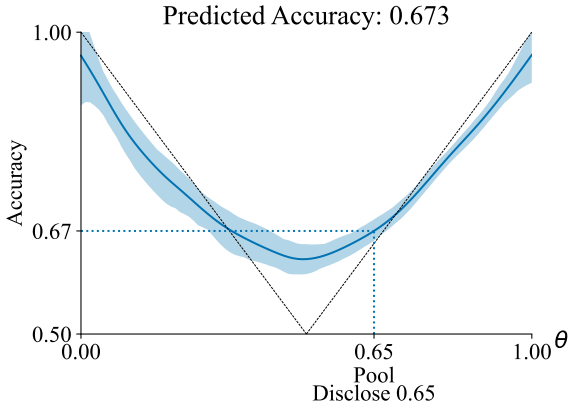
(a) Full Disclosure + Automation (FDA)



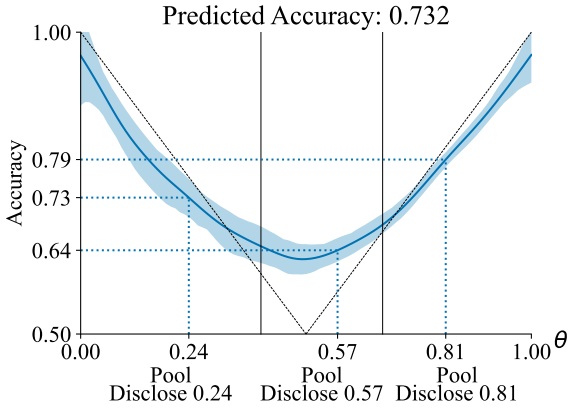
(b) No Disclosure + Automation (NDA)



(c) No Disclosure + No Automation (NDNA)



(d) Stoplight + No Automation (SL)



Note: Policies tested in the Stage 2 experiment. The function V is estimated using local linear regression from Stage 1 data. The bandwidth is chosen via leave-one-out cross-validation to minimize mean squared error. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level. The dashed lines indicate the accuracy under automation of $\max\{\theta, 1-\theta\}$. The dotted lines indicate the assessments disclosed to participants and the associated accuracy predicted by V.

points lower than the optimal policy (FDA), the predicted value of direct human-AI collaboration is very small.

The intuition for why predicted accuracy under FDA or NDA is almost identical is that V is relatively flat on the intervals of non-automated cases, $[0.33, 0.69]$ (for FDA) or $[0.40, 0.67]$ (for NDA). Because the benefit of disclosing information comes from the convexity of V, this implies that the benefit of disclosing AI assessments on this interval of cases is small.

Note that it is optimal to automate cases with a wider range of AI assessments under

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

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

(c) **Full Disclosure + No Automation** (**FDNA**): This is the optimal policy without automation (i.e., the solution to problem (2)), which is also the policy used in Stage 1. The predicted accuracy of this policy equals the average accuracy in Stage 1, 73.5%.

(d) **No Disclosure + No Automation** (**NDNA**): With no disclosure or automation, participants are only informed of the share of true cases in the database, which is 65.4%. The predicted accuracy of this policy is 67.3%.

(e) **Stoplight** (**SL**): The final policy we consider approximates full disclosure using a very simple signal distribution. Specifically, we calculate the optimal partition of AI assessments into K intervals and disclose the average AI assessment within each interval:

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

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

We consider "Stoplight" with $K = 3$ for two reasons. First, predicted accuracy with $K = 3$ is 73.2%, which we expect to be indistinguishable from the predicted accuracy of 73.5% when $K = \infty$.[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 a system in which the AI reports only that each case is

[28] Predicted accuracy for other values of K are shown in Figure B.4.

either "Likely False/Red," "Uncertain/Yellow," or "Likely True/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] The predicted accuracy at these three assessments is 0.73, 0.64, and 0.79 respectively.

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

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

2. **Human information is valuable.** Accuracy under the optimal policy with automation (75.1%) exceeds that under full automation (73.3%).

3. **Human-AI collaboration does not outperform selective automation.** Accuracy under the optimal policy with automation (75.1%) does not significantly exceed that under the optimal no-collaboration policy (74.8% under NDA).

4. **AI assistance is valuable when automation is infeasible.** Accuracy under the optimal policy without automation (73.5%) exceeds that without AI assistance (67.3% under NDNA).

5. **Simple disclosure polices are approximately optimal.** Accuracy under SL (73.2%) approximates that under the optimal policy without automation (73.5%).

From the perspective of testing Assumption 1, it is worth highlighting that the quantitative predictions for these policies are counterfactual (except for FDNA). The no-disclosure and Stoplight policies also provide counterfactual AI assessments for a given statement. The accuracy predictions for these policies are thus particularly demanding tests of our framework. The minimum detectable effect size in the second stage is 1.4 percentage points at a 5% significance level for 80% power. The experiment is thus powered to detect the predicted differences in points 1, 2 and 4 and to rule out large differences in points 3 and 5.

4.4 Restrictions on the Design Space

Three restrictions on our design space merit discussion. First, while we let the AI flexibly disclose information to human decision-makers, we do not consider systems that elicit

[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 Release Conditions Matrix.

[30] It is a coincidence that the middle interval under Stoplight coincides with Θ^{aut} under NDA.

humans' assessments and combine them with the AI's information. That is, we consider "one-way" communication from AI to humans, not "two-way" communication. An analysis of communication from humans to AI should consider the possibility that humans make strategic reports, which would require a different design. However, in our setting, one-way communication turns out to be without loss of optimality: Section 5.2 shows that the maximum accuracy attainable with access to both human and AI assessments under FDNA is indistinguishable from that under FDA (the optimal policy without elicitation).

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

Third, we do not tailor the policy to predictable heterogeneity across participants, although the sufficient statistic V is a predictable function of baseline comprehension questions, effort, or accuracy on initial statements (Figure D.6). However, Table D.11 shows that tailoring policies to this heterogeneity does not substantially improve predicted accuracy.[31]

5 Stage 2 Results

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

We estimate the average outcome for each policy in Stage 2 using the regression:

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

where y_{ij} is an outcome for statement i by participant j, and β_k is the average outcome under policy k.[32] We cluster standard errors to allow for $Cov(\varepsilon_{ij}, \varepsilon_{i'j'}) \neq 0$ if either $i' = i$ or $j' = j$, but set it to zero otherwise. Estimated treatment effects relative to FDNA are therefore given by $\beta_k - \beta_{k_0}$, where k_0 is the baseline FDNA policy.

Summary statistics analogous to Table 1 comparing FDNA in the two stages are presented in Table B.10. The main difference, aside from sample size, is that average performance is

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

[32] Participants only assess non-automated cases. The dependent variables of interest are system accuracy and effort. We therefore use 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 automated AI assessments under a given policy, and $\bar{y}$ is the average outcome among automated cases. For accuracy, $\bar{y} = \mathbb{E}\left[\max\{\theta, 1-\theta\} | \theta \in \Theta^{\text{Aut}}\right]$; for effort measures, $\bar{y} = 0$.

slightly lower in Stage 2. We discuss this difference further below.

5.1 Validity of the Sufficient-Statistic Approach

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

The experiment confirms all of our qualitative predictions:

1. **Automation is valuable.** Accuracy under FDA exceeds that under FDNA. The estimated difference is 2.6 percentage points ($p < 0.001$). The predicted difference is 2.5 and 1.7 percentage points using the Stage 2 and Stage 1 estimates of V, respectively.

2. **Human information is valuable.** Accuracy under FDA exceeds that under full automation. The estimated difference is 1.6 percentage points ($p < 0.001$). The predicted difference is 1.5 and 1.8 percentage points using Stage 2 and Stage 1 estimates.

3. **Human-AI collaboration does not outperform selective automation.** Accuracy under FDA does not significantly exceed that under NDA ($p = 0.44$). Human-AI collaboration increases accuracy by 0.2 percentage points, which is within 0.2 percentage points of our prediction using either estimate of V.

4. **AI assistance is valuable when automation is infeasible.** Accuracy under FDNA exceeds that under NDNA. The estimated difference is 3.5 percentage points ($p < 0.001$). The predicted difference is 5.5 and 6.2 percentage points using Stage 2 and Stage 1 estimates.

5. **Simple disclosure policies are approximately optimal.** Accuracy under FDNA does not significantly exceed that under SL ($p = 0.724$). The experimental estimates suggest a small gain of 0.2 percentage points from using SL over FDNA, whereas the predictions suggest a loss of 0.3 percentage points using either estimate of V.

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

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

Table 2: Estimated Versus Predicted Accuracy

		Stage 2		Stage 1	
Treatment	Stage 2 Estimate	Predicted	P-value	Predicted	P-value
	(1)	(2)	(3)	(4)	(5)
Panel A:					
Full Disclosure + No Automation	0.723	-	-	0.735	0.014
(FDNA)	(0.004)			(0.003)	
Panel B: Automation					
Full Disclosure	0.749	0.748	0.781	0.751	0.342
(FDA)	(0.002)	(0.003)		(0.002)	
No Disclosure	0.747	0.744	0.345	0.748	0.734
(NDA)	(0.001)	(0.003)		(0.002)	
Panel C: No Automation					
No Disclosure	0.689	0.669	0.035	0.673	0.022
(NDNA)	(0.004)	(0.009)		(0.006)	
Stoplight	0.725	0.720	0.484	0.732	0.175
(SL)	(0.004)	(0.006)		(0.004)	
Joint Test	–	–	0.170	–	0.010

Note: Column (1) is the estimated accuracy from Stage 2 data. Column (2) is the predicted accuracy computed from the Stage 2 V estimate. Column (4) is the predicted accuracy computed from the Stage 1 V estimate, except for the FDNA row, which contains the observed accuracy in Stage 1. Columns (3) and (5) contain the p-value from a test of the null hypothesis that the Predicted and Estimated values are equal. Standard errors are in parentheses. Predicted standard errors are computed via block bootstrap clustered at the participant level, and the Stage 2 Estimate standard errors are two-way clustered at the participant and case level. The p-values are based on a block bootstrap clustered at the participant level.

There are two departures from the model's quantitative predictions. First, we correctly predict the quantitative value of automation (prediction 1) relative to the Stage 2 estimate of V but not relative to the Stage 1 estimate; and second, we mispredict the quantitative value of AI assistance when automation is infeasible (prediction 4) using V from either stage.

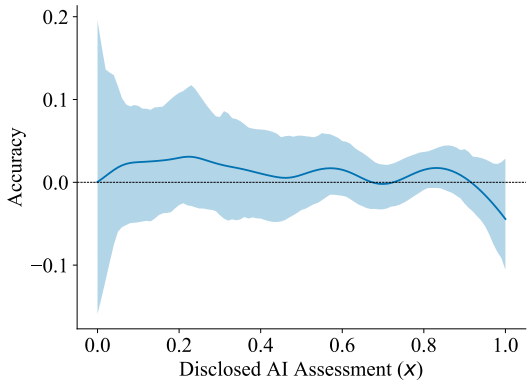
Table 2 provides results on the specific policies where our predictions do not match the experimental estimates. Columns 3 and 5 show that we cannot reject that accuracy under FDA, NDA, or SL equals the predicted accuracy using either estimate of V. However, accuracy under FDNA is 1.2 percentage points lower ($p = 0.01$) than accuracy from Stage 1; and accuracy under NDNA is 2.0 and 1.6 percentage points higher than the predicted accuracy using the estimate of V from Stage 2 and Stage 1, respectively (p-values < 0.05).

There are two distinct reasons why the predictions from the two stages may miss the experimental estimates. The first is that details of the experimental protocol might affect performance. For instance, there may be subtle differences in participants, participants may learn to use the AI differently, and there may be effects of exposure to multiple treatments in

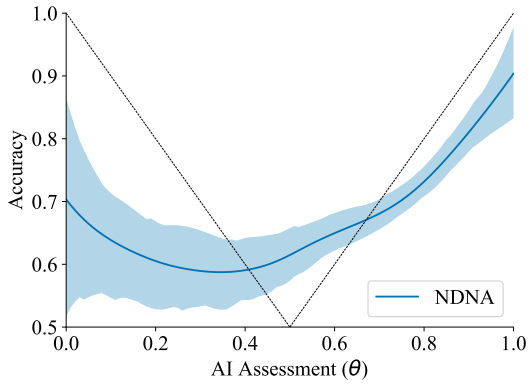
Stage 2. Figure 8a shows that the estimates of V from the two stages are similar—we cannot reject that the function V estimated in the two stages using data from FDNA are the same ($p = 0.28$). However, Panel A of Table 2 shows that accuracy under FDNA is lower in Stage 2 (72.3%) than in Stage 1 (73.5%).[33] These differences do not speak to Assumption 1, but they may explain why the predicted value of automation based on the Stage 1 estimate of V is quantitatively inaccurate, while the Stage 2 prediction is accurate.

Figure 8: Stability

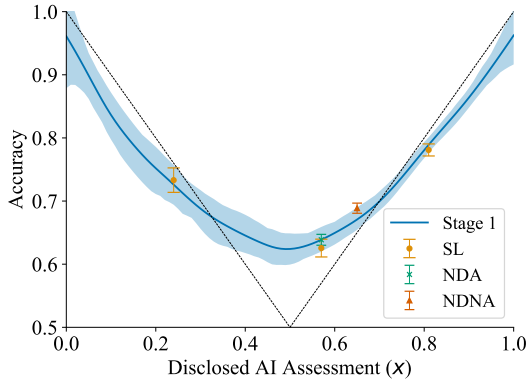
(a) Difference in V between Stage 1 and 2



(b) Accuracy as a Function of θ under NDNA



(c) Predicted vs Estimated Accuracy



Note: Figure 8a plots the difference in accuracy between Stage 1 and Stage 2, $V_1(x) - V_2(x)$, estimated by local linear regression for each stage. Figure 8b plots accuracy conditional on θ for Stage 2 under NDNA estimated by local linear regression. Figure 8c plots accuracy conditional on x for Stage 1 via the function V estimated by local linear regression. The dashed lines indicate the accuracy under automation of $\max\{x, 1-x\}$. The accuracy by x under SL, NDA, and NDNA is estimated by regressing an indicator for correct classification on indicators for each AI assessment shown. The 95% point-wise confidence intervals for the SL, NDA, and NDNA estimates in Figure 8c are two-way clustered at the participant and case level. For all three figures, the 95% uniform confidence band is computed via bootstrap accounting for clustering at the participant and case level.

[33]The difference is statistically significant ($p = 0.01$). Participants were also faster in Stage 2. One explanation is that participants assessed more statements in this stage. Figure A.2 shows that participants become faster—and possibly more accurate—in Stage 1 but not Stage 2 after assessing about 25 statements.

The second reason is a violation of Assumption 1: participants' accuracy may not depend only on the mean AI assessment. This can explain the greater accuracy under NDNA in Stage 2 relative to the prediction. Specifically, a likely explanation is that cases where the AI is confident are also easier for human participants, so that participants' average accuracy under NDNA is higher than their accuracy on cases where the AI assessment is uninformative (cases where $\theta = \phi$). Figure 8b points to this hypothesis. It plots participant accuracy as a function of the AI assessment θ under NDNA, where θ is *not* disclosed. Assumption 1 implies that this curve must be linear in θ for a Bayesian decision-maker with conditionally independent private information, which Figure 8b contradicts. That said, the magnitude of the violation is small: participants' average accuracy under NDNA is only 1.6 percentage points higher than their accuracy on cases with the average AI assessment ϕ, suggesting that cases where the AI is more confident are only slightly easier for human participants.[34]

Even if we see a small violation of Assumption 1, the policies we derived under Assumption 1 seem very likely to be exactly optimal in our setting. Specifically, nothing in our data suggests a violation of the weaker Assumption 1' with a convex $\tilde{V}$, and FDA remains optimal under these assumptions (and FDNA remains the optimal no-automation policy).[35]

Another demanding test of Assumption 1 compares the predicted accuracy at specific counterfactual posteriors x to estimates from Stage 2. Figure 8c displays these estimates, which show the realized Stage 2 accuracy at the induced posteriors under NDA and NDNA, as well as at each of the three induced posteriors under SL. The predictions from NDA and SL match the corresponding values of $V(x)$ from the Stage 1 estimate of V.[36] As previously mentioned, the prediction for NDNA slightly misses. As a benchmark, the algorithmic triage approach would use data from NDNA to predict accuracies of 0.61, 0.64 and 0.74 for the three intervals in SL. These predictions would miss the estimates from our experiment of 0.73 and 0.78 for the first and third intervals ($p < 0.01$). In contrast, our Stage 1 predictions of 0.73 and 0.79 are much closer and are statistically indistinguishable from the Stage 2 estimates (p-values = 0.32, 0.89). Thus, our accuracy predictions are validated for specific counterfactually induced posterior assessments, not just on average.

Overall, although we slightly mispredict accuracy in one treatment, the sufficient-statistic approach provides a useful guide to designing automation/disclosure policies in our setting.

[34]To benchmark this number, note that predicted accuracy under the opposite assumption that human and AI signals are perfectly correlated is 73.5%—i.e., the same prediction as under FDNA—which exceeds actual accuracy under NDNA by a much larger 4.6 percentage points.

[35]For example, the data suggest that cases where the AI is confident are also easier for humans, so we would expect that $\tilde{V}(.5 \cdot \delta_0 + .5 \cdot \delta_1) > \tilde{V}(\delta_{.5})$, a violation of Assumption 1. But we would certainly also expect that $.5\tilde{V}(\delta_0) + .5\tilde{V}(\delta_1) > \tilde{V}(.5 \cdot \delta_0 + .5 \cdot \delta_1)$, consistent with Assumption 1' and convexity of $\tilde{V}$.

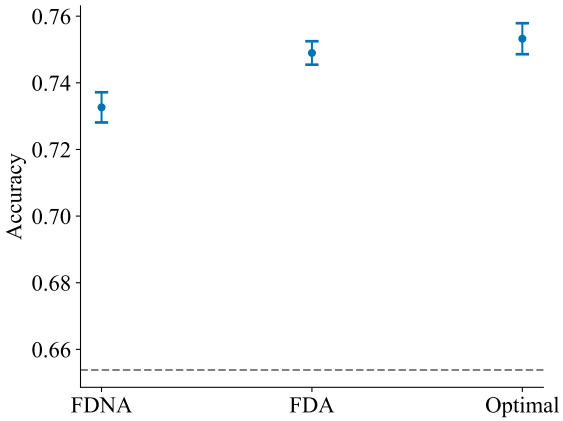
[36]We do not reject the joint test that the estimated Stage 2 accuracy at each induced posterior equals the predicted accuracy (p =0.362).

5.2 Optimal Classifier Benchmark

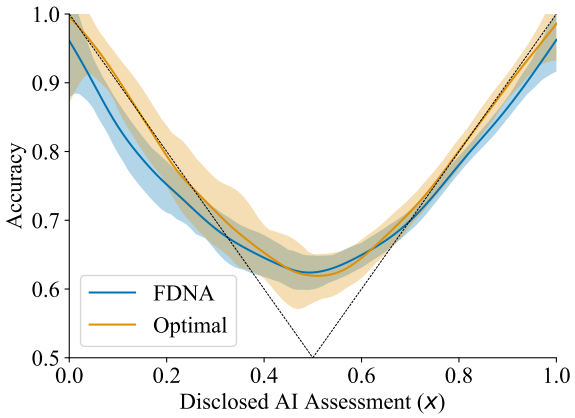
We now calculate the accuracy of an optimal classifier V^{Opt} with access to both participants' reported assessments p_{ij} under FDNA and the AI assessments θ_i.[37] The optimal classification for case i is True if and only if $\Pr(\omega = 1|p_{ij}, \theta_i)$ exceeds 0.5. We non-parametrically estimate $\Pr(\omega = 1|p, \theta)$ as a function of (p, θ) using the FDNA sample from both stages. To avoid overfitting, we use a penalized logistic regression with polynomial terms in p and θ with cross-validation for model selection (see Appendix G for details). Figure 9a compares accuracy under the tested policies to the optimal classifier.

Figure 9: Comparing Participant and Optimal Classifier Accuracy

(a) Participant vs. Classifier Accuracy



(b) Participant vs. Classifier V



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

The first result from this exercise is that average accuracy under the optimal classifier (75.3%) is approximately equal to accuracy under FDA (74.9%). This result implies that one-way communication from AI to humans is without loss of optimality: the optimal policy in our design space without elicitation of participants' assessments (FDA) cannot be significantly improved by elicitation. Intuitively, this is a consequence of two properties of the function V^{Opt} with the optimal classifier, which is plotted alongside V with human decision-makers in Figure 9b. First, V^{Opt} is indistinguishable from human accuracy V for AI assessments where delegation to a human is optimal (i.e., where $V(\theta) \geq \max\{\theta, 1-\theta\}$). Second, $V^{\text{Opt}}(\theta)$ is indistinguishable from AI accuracy $\max\{\theta, 1-\theta\}$ for AI assessments where automation is optimal (i.e., where $V(\theta) < \max\{\theta, 1-\theta\}$). Together, these properties imply that selective automation achieves the optimal classifier benchmark.

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

A second result is that accuracy under the optimal classifier is significantly greater than that under FDNA. This implies a substantial impact of participants' under-response to AI information on accuracy: if participants were correctly-specified Bayesians, their accuracy would be at least as high as the optimal classifier's, because they know the AI assessment and may also have additional information. This comparison gives a lower bound for the impact of non-Bayesian updating on participant accuracy—at least $(75.32\% - 73.26\%)/(1 - 73.26\%) = 7.7\%$ of incorrect classifications under FDNA are attributable to deviations from Bayesian updating. Section 6 unpacks the deviations that are responsible for this result.

5.3 Impact on Effort

Table 3 presents estimated treatment effects on our three measures of participant effort, relative to the baseline FDNA policy. It uses estimates from the model in equation (5) and reports β_0 for FDNA and $\beta_k - \beta_0$ for the remaining policies.

Disclosing AI assessments crowds out human effort, consistent with Figure 6b: our effort measures are between 9% and 11% lower under FDNA relative to NDNA. While this effect is substantial, it is smaller than some estimates in the literature: for example, Dell'Acqua (2022) finds that disclosing more precise AI assessments reduced effort by nearly 40%.

As with the estimated treatment effects on accuracy, this result is robust to the variations described earlier (see Tables A.5, A.6, A.7, and A.8).[38]

6 Overconfidence, AI Neglect, and Effort Crowd-Out

The estimates in Sections 4 and 5 show that our participants under-respond to AI assessments and reduce effort when presented with confident AI assessments. This section analyzes participants' biases in belief updating and the impact of effort crowd-out on accuracy.[39]

We distinguish between participants' *overconfidence* in the precision of their own information and *under-confidence* in the precision of AI information—which we refer to as *AI neglect*. Empirically distinguishing overconfidence from AI neglect requires additional assumptions to identify the distribution of participants' private information and their model of belief updating. Under these assumptions, we also show that the reduction in participant accuracy due to measured effort crowd-out is modest in magnitude.

[38]When using the across-participant design based on the first treatment encountered (Table A.5), the treatment effects are similar to the within-participant design, except the baseline effort measures are higher, probably due to learning and becoming faster over time (see Figure A.2). However, learning does not materially impact V, and the figure shows that the impact of learning on accuracy is minimal.

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

Table 3: Average Treatment Effects on Effort

Treatment	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)
Panel A: No Automation Baseline (β_0)			
Full Disclosure	0.630	0.372	44.551
(FDNA)	(0.009)	(0.009)	(0.730)
Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)			
Full Disclosure	-0.357	-0.209	-24.515
(FDA)	(0.007)	(0.007)	(0.560)
No Disclosure	-0.412	-0.240	-28.523
(NDA)	(0.007)	(0.007)	(0.586)
Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)			
No Disclosure	0.064	0.046	3.749
(NDNA)	(0.006)	(0.006)	(0.586)
Stoplight	0.003	0.001	0.091
(SL)	(0.005)	(0.006)	(0.529)
Observations	80000	80000	80000

Note: Average treatment effects (estimated using equation 5) of different polices on effort. In FDA and NDA, outcomes are adjusted to account for automation (Footnote 32). Standard errors are two-way clustered at the participant and case level.

6.1 Over- and Under-Inference

We define overconfidence and AI neglect using a general definition of over- or under-inference from a signal based on the following model. 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 $s \in \mathbb{R}$. For example, this property holds if each signal s_n is calibrated (i.e., $s_n \in [0, 1]$ and $\Pr(\omega = 1|s_n = s) = s$ for all $s \in [0, 1]$).

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

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

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

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

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

This definition of over-inference is novel as far as we know, although it has predecessors in the non-Bayesian updating literature. The closest is in Augenblick et al. (2025), which defines a notion of the perceived strength $\hat{\mathbb{S}}(s)$ of a signal s and says 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 based on the model in Grether (1980). In particular, if signals are conditionally independent and calibrated, and we rewrite (6) in an equivalent form where the derivatives are taken with respect to 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.[40]

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 $p(s, x)$. Applying the definition above, humans are *overconfident* in their own signal if they over-infer from s, and 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 the human signal to be calibrated, so that $s_{ij} = P(\omega_i = 1 | s_{ij})$.*

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

Assumption 2.1 imposes two restrictions. First, the distribution of human signals does

[40] In the conditionally independent case, the models in Grether (1980) and Agarwal et al. (2023) assume that $\text{logit}\, p(\mathbf{s}) = \sum_{n=1}^{N} a_n \left(\text{logit Pr}(\omega = 1 | s_n) - \text{logit Pr}(\omega = 1)\right) + b \text{ logit } Pr(\omega = 1)$ for parameters $a_1, \ldots, a_N, b$. For $p^{\text{Bayes}}(\cdot)$, Bayes' rule gives $a_1 = \ldots = a_N = b = 1$.

not depend on the disclosure policy or the disclosed AI signal x_i conditional on $\omega_i, e_{ij}, \theta_i$.[41] In particular, our observed measures of effort e_{ij}—time taken, an indicator for the reported use of external sources, and an indicator for clicking the Google search link—are sufficient controls for the dependence of the human signal s_{ij} on the disclosed AI signal x_i. Second, while the distribution of effort can vary across human participants, the signal distribution is the same across participants conditional on effort.

Assumption 2.2 imposes three restrictions. First, the human assessment p_{ij} depends only on the human signal s_{ij} and the disclosed AI assessment x_i. Second, the assessment is monotone in the human signal.[42] For example, Assumption 2.2 holds if humans are Bayesian with conditionally independent signals. It also holds if humans are quasi-Bayesians who act as if their signals are conditionally independent of the AI signal and may over- or under-weight either signal, as in Grether (1980). Third, the function $p(\cdot)$ is the same for all participants, as we estimate a single updating rule rather than attempting to distinguish heterogeneous updating rules across participants.[43]

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

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

The conditional CDF $F_{p|\theta}$ is observed under FDNA and we estimate it non-parametrically.[44] The remaining task is to identify and estimate $F_{s|\theta}$. We accomplish this in two steps. First, we identify and estimate the human signal distribution $G_{\omega_i, e_{ij}, \theta_i}$ using data under NDNA. By Assumption 2.1, $G_{\omega_i, e_{ij}, \theta_i}$ is independent of the disclosure policy and the disclosed AI assessment x_i, conditional on $(\omega_i, e_{ij}, \theta_i)$. Under NDNA, the disclosed AI assessment x_i is constant at the prior $\phi = \Pr(\omega = 1)$ but we observe continuous assessments $p_{ij} = p(s_{ij}, x_i)$. Since x_i is constant, Assumption 2.2 implies that p_{ij} is a monotonic function of only s_{ij}, and hence $\Pr(\omega_i = 1|p_{ij}) = \Pr(\omega_i = 1|s_{ij}) = s_{ij}$. Thus, under NDNA, s_{ij}, ω_i, e_{ij}, and θ_i are observable and we can identify and estimate $G_{\omega_i, e_{ij}, \theta_i}$ non-parametrically (see footnote 44).

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

[42] Identification does not depend on the assessment being monotone in the disclosed AI assessment x_i.

[43] This follows prior work estimating the Grether (1980) model (Benjamin, 2019).

[44] We estimate all conditional CDFs of the 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 x is a vector. We estimate this for a grid of z to trace out the conditional CDF (Chernozhukov et al., 2013). If the estimates are non-monotonic, we apply the rearrangement procedure described in Chernozhukov et al. (2010).

Next, the conditional CDF $F_{s|\theta}$ can be calculated from $G_{\omega_i,e_{ij},\theta_i}$—identified from the NDNA data—by integrating over the observed joint distribution of ω_i and e_{ij} under FDNA. We estimate this distribution by fitting a conditional distribution model to 100,000 simulated draws from the joint distribution of s_{ij}, e_{ij}, θ_i, and ω_i under FDNA. We simulate these draws from the joint distribution of e_{ij}, θ_i, and ω_i using an accept/reject sampler.[45,46] We then sample s_{ij} from the conditional distribution $G_{\omega_i,e_{ij},\theta_i}$ using inverse transform sampling.

Finally, we use a plug-in estimator that replaces the conditional distributions of p and s with the estimated analogues in equation (7).[47]

6.3 Overconfidence or AI Neglect?

We can now compare the estimated belief updating rules $p(s,x)$ to the Bayesian benchmark $p^{\text{Bayes}}(s,x)$ to decompose the AI under-response found in Section 4 into overconfidence and AI neglect. We estimate $p^{\text{Bayes}}(s,x)$ through a penalized logistic regression of ω on s and x in the 100,000 samples of ω, s, and θ described above (see Appendix G for details).

Figure 10 presents estimates of our participants' update function p (blue curve) and the Bayesian benchmark rule p^{Bayes} (orange curve), as well as the Bayesian benchmark imposing conditional independence (green curve; a line of slope 1 in log odds space). The panels hold either s or x fixed at a specific value while varying the other signal in log odds space.[48] A first observation is that the two Bayesian benchmarks are quite similar, implying that conditional independence ($s \perp x|\omega$) is a good approximation.

Figures 10a–10c show strong evidence of overconfidence. Recall that Section 6.1 defines overconfidence as logit p being steeper than logit p^{Bayes} when s varies, and AI neglect as logit p being flatter than logit p^{Bayes} when x varies. Correspondingly, the slope of logit p with respect to s in Figures 10a–10c is much larger than the Bayesian benchmark. This overconfidence results in participants reporting more extreme probability assessments than a calibrated decision maker (as shown in Figure 6a), as well as in AI under-response.

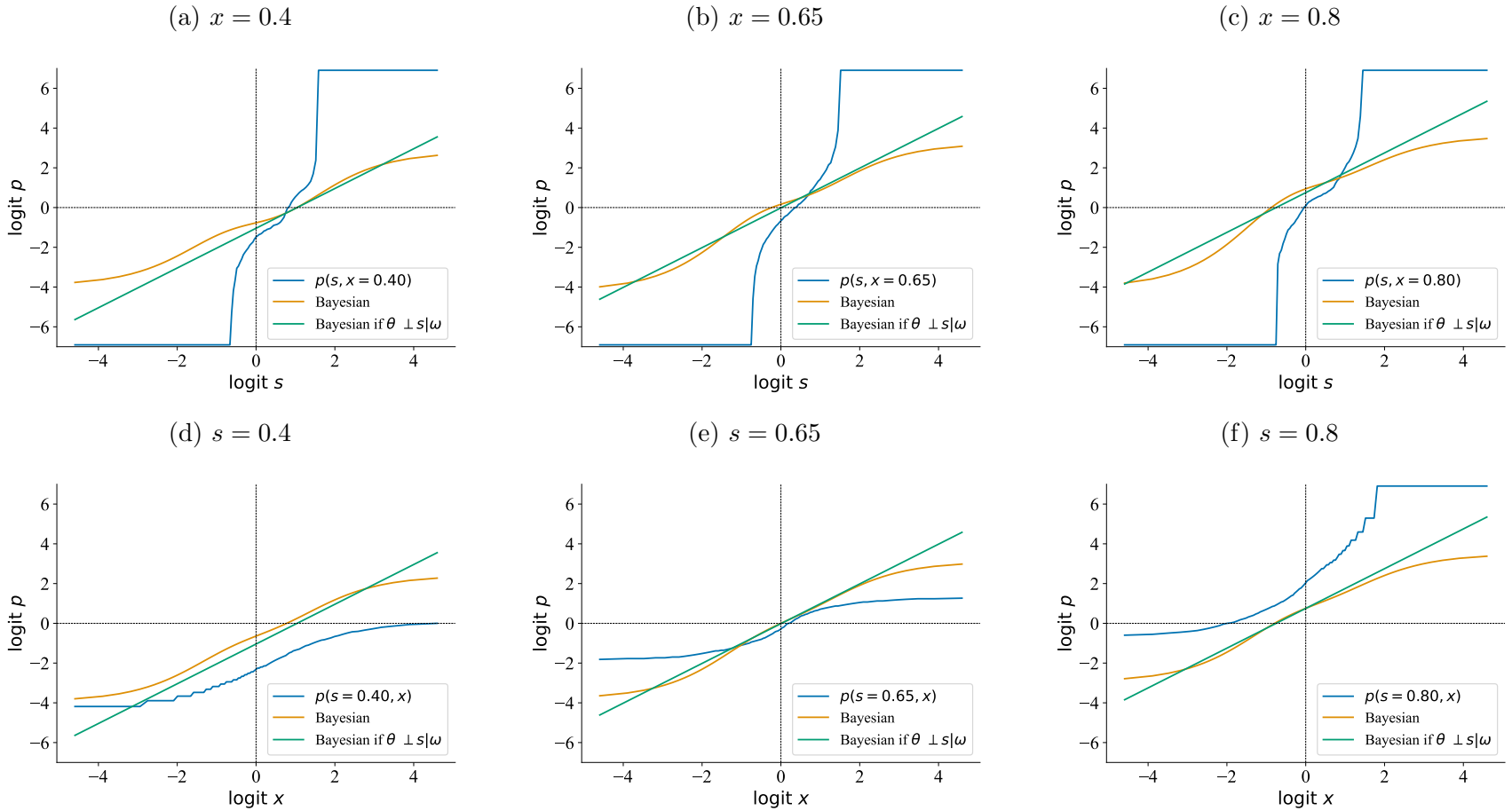
[45] We estimate the joint distribution of e_{ij}, θ_i, and ω_i under FDNA using a kernel density estimator. We use a Gaussian kernel for all continuous variables and Silverman's rule to select bandwidths (Silverman, 2018) and a bandwidth of 0 for all binary variables.

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

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

[48] Appendix Figure C.5 presents a corresponding surface plot.

Figure 10: Human vs. Bayesian Update Rule



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

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

Overall, we find strong evidence of overconfidence and some evidence of AI neglect.

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

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

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

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

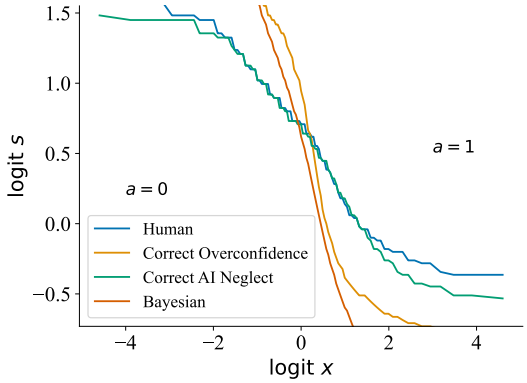
Figure 11a plots the decision threshold in $(\text{logit}\,x, \text{logit}\,s)$-space for humans, Bayesians, and humans corrected for overconfidence or AI neglect. The decision threshold for overconfidence-corrected humans is close to the Bayesian benchmark, while the threshold for AI neglect-corrected humans is close to that for 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 for the Bayesian benchmark). Thus, overconfidence—not AI neglect—is the main reason our participants deviate from optimal Bayesian decisions.

Our result that AI under-response is primarily due to overconfidence rather than AI neglect differs from that in Agarwal et al. (2023), which finds evidence for AI neglect but not overconfidence among professional radiologists.[49] One hypothesis for this difference is that professional decision-makers such as radiologists tend to understand their own abilities but distrust outside advice, while amateurs such as our participants over-estimate their own abilities but are more open to advice.

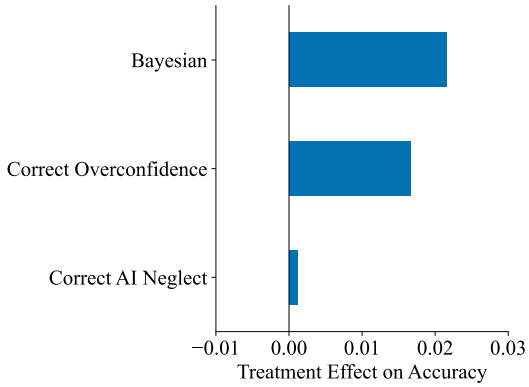
[49] Agarwal et al. (2023) estimates the Grether model where $\text{logit}\,p(s,x) = b\,\text{logit}\,x + c\,\text{logit}\,s$, finding that $b = 0.26$ and $c = 0.87$. In contrast, estimating the same model with our data yields $b = 0.8$ and $c = 1.6$.

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

6.4 Impact of Effort Crowd-Out on Human Signal Quality

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

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

Panel B shows that this effort crowding-out also reduces three measures of human signal precision—effort crowding-out increases the root mean-square error of the human signal; reduces the probability that the human signal alone would result in a correct classification; and increases the probability that the human signal does not overturn the prior favoring classifying cases as True. All of these reductions in precision are concentrated on statements that the AI is confident are false ($x < 0.25$). A possible explanation for the asymmetry between cases where $x < 0.25$ and where $x > 0.75$ is that, since cases where $x < 0.25$ are rare (see Figure 4), disclosing that $x < 0.25$ has a larger effect on participant beliefs.

Overall, Table 4 provides modest evidence that effort crowding-out due to AI disclosure reduces human signal precision and contributes to the value of selective automation.

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

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

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

7 Conclusion

Collaboration between humans and AI can profoundly affect organizational decision-making and job design, and its importance will only grow over time (Daugherty and Wilson, 2018; Mollick, 2024). The design of human-AI collaborative systems is thus a pressing concern. The standard approach to this problem in the literature is "algorithmic triage" (Raghu et al., 2019; Mozannar and Sontag, 2020; Agarwal et al., 2023), which decides which cases to automate and which to assign to humans, with or without AI assistance. However, this approach does not allow richer designs that partially disclose AI information, and it does not account for the endogenous response of human beliefs and effort to automation and AI disclosure policies. Moreover, the dimensionality of the design space and the complexity of human responses frustrate the search for an optimal design through direct experimentation.

We contribute a method for finding the optimal human-AI collaborative design in binary classification problems by estimating a simple sufficient statistic: the probability of correct classification as a function of the disclosed posterior. We validate this approach with an online fact-checking experiment, where we find that the optimal policy automates cases where the AI is confident and delegates the remaining cases to human decision-makers while fully disclosing

the AI assessment. At the same time, even simpler policies—such as selective automation without direct human-AI communication—are approximately optimal in our setting. The value of automation stems from our participants' under-response to AI information, which in turn results from over-confidence in the precision of their own information, rather than under-confidence in the AI.

One avenue for future research is to enlarge the space of collaborative policies considered. For example, while we document substantial effort response to AI information, we do not consider the joint design of an information disclosure policy and an incentive contract. Similarly, we document significant biases in belief updating in response to AI information, but we do not consider policies targeted at reducing these biases, such as training humans to put appropriate weights on AI information and their own. In addition, while our optimal classifier analysis suggests that any benefits from eliciting humans' assessments upfront are negligible, we do not consider more complex dynamic communication protocols where humans and AI communicate in multiple rounds. Such protocols would raise new issues such as the possibility that humans make strategic reports to the AI.

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

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

Finally, another avenue for research is considering richer information structures where the human and AI signals are not conditionally independent, violating Assumption 1. In this case, disclosing the AI's posterior is not sufficient, as classification accuracy can depend on further details of the AI's information. This possibility relates to how AI predictions or recommendations should be explained to human decision-makers, as well as to additional gains from unstructured human-AI communication (e.g., via an LLM). While some studies find that AI explanations have a small overall impact on human accuracy in classification problems (Green and Chen, 2019; Bansal et al., 2021), there is little work on the form of AI explanations and how they relate to dependencies between human and AI information. These issues may play a role in determining which settings feature a larger scope for direct human-AI collaboration and which (like ours) do not.

References

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

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Harvard Business Press, April 2018.

Allen, Jennifer, Antonio A Arechar, Gordon Pennycook, and David G Rand, "Scaling up fact-checking using the wisdom of crowds," *Science Advances*, 2021, *7* (36).

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," *AEA Papers and Proceedings*, May 2020, *110*, 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.

Bansal, Gagan, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in "Proceedings of the 2021 CHI conference on human factors in computing systems" 2021, pp. 1–16.

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

Blackwell, David, "Equivalent Comparisons of Experiments," *The Annals of Mathematical Statistics*, 1953, *24* (2), 265–272. https://www.jstor.org/stable/2236332.

Brynjolfsson, Erik, Danielle Li, and Lindsey Raymond, "Generative AI at work," *The Quarterly Journal of Economics*, 2025, *140* (2), 889–942.

Carrell, Scott E, Bruce I Sacerdote, and James E West, "From natural variation to optimal policy? The importance of endogenous peer group formation," *Econometrica*,

2013, *81* (3), 855–882.

Chen, Daniel L, Martin Schonger, and Chris Wickens, "oTree—An open-source platform for laboratory, online, and field experiments," *Journal of Behavioral and Experimental Finance*, 2016, *9*, 88–97.

Chernozhukov, Victor, Iván Fernández-Val, and Alfred Galichon, "Quantile and probability curves without crossing," *Econometrica*, 2010, *78* (3), 1093–1125.

_ , _ , and Blaise Melly, "Inference on counterfactual distributions," *Econometrica*, 2013, *81* (6), 2205–2268.

Chetty, Raj, "Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods," *Annu. Rev. Econ.*, 2009, *1* (1), 451–488.

Clippel, Geoffroy De and Xu Zhang, "Non-bayesian persuasion," *Journal of Political Economy*, 2022, *130* (10), 2594–2642.

Daugherty, Paul R and H James Wilson, *Human+ machine: Reimagining work in the age of AI*, Harvard Business Press, 2018.

Dell'Acqua, Fabrizio, "Falling asleep at the wheel: Human/AI Collaboration in a Field Experiment on HR Recruiters," Technical Report, Working Paper 2022.

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

Dreyfuss, Bnaya and Ruru Hoong, "Calibrated Coarsening: Designing Information for AI-Assisted Decisions," *Working Paper*, 2025.

Dubé, Jean-Pierre and Sanjog Misra, "Personalized pricing and consumer welfare," *Journal of Political Economy*, 2023, *131* (1), 131–189.

Dworczak, Piotr and Anton Kolotilin, "The Persuasion Duality," *Theoretical Economics*, 2024, *19* (4), 1701–1755.

_ and Giorgio Martini, "The simple economics of optimal persuasion," *Journal of Political Economy*, 2019, *127* (5), 1993–2048.

Fang, Zheng and Juwon Seo, "A projection framework for testing shape restrictions that form convex cones," *Econometrica*, 2021, *89* (5), 2439–2458.

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, "A Rothschild-Stiglitz approach to Bayesian persuasion," *American Economic Review*, 2016, *106* (5), 597–601.

Green, Ben and Yiling Chen, "The principles and limits of algorithm-in-the-loop decision making," *Proceedings of the ACM on human-computer interaction*, 2019, *3* (CSCW), 1–24.

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

Guo, Zhijiang, Michael Schlichtkrull, and Andreas Vlachos, "A survey on automated fact-checking," *Transactions of the Association for Computational Linguistics*, 2022, *10*, 178–206.

Guo, Ziyang, Yifan Wu, Jason Hartline, and Jessica Hullman, "The Value of Information in Human-AI Decision-making," *arXiv preprint arXiv:2502.06152*, 2025.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Springer Series in Statistics, 2 ed., New York: Springer, 2009.

Hirano, Keisuke, Guido W Imbens, and Geert Ridder, "Efficient estimation of average treatment effects using the estimated propensity score," *Econometrica*, 2003, *71* (4), 1161–1189.

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

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

Kamenica, Emir and Matthew Gentzkow, "Bayesian Persuasion," *American Economic Review*, October 2011, *101* (6), 2590–2615.

Kleinberg, Jon, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan, "Human Decisions and Machine Predictions," *The Quarterly Journal of Economics*, August 2017, *133* (1), 237–293.

Kolotilin, Anton, "Optimal information disclosure: A linear programming approach," *Theoretical Economics*, 2018, *13* (2), 607–635.

_ , Tymofiy Mylovanov, Andriy Zapechelnyuk, and Ming Li, "Persuasion of a privately informed receiver," *Econometrica*, 2017, *85* (6), 1949–1964.

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," *arXiv preprint arXiv:2112.11471*, 2021.

Li, Danielle, Lindsey R Raymond, and Peter Bergman, "Hiring as exploration," Technical Report, National Bureau of Economic Research 2020.

Lowenkamp, Christopher T, "The development of an actuarial risk assessment instrument for US Pretrial Services," *Federal Probation*, 2009, *73*, 33.

McLaughlin, Bryce and Jann Spiess, "Designing Algorithmic Recommendations to Achieve Human-AI Complementarity," *arXiv preprint arXiv:2405.01484*, 2024.

Misra, Sanjog and Harikesh S Nair, "A structural model of sales-force compensation dynamics: Estimation and field implementation," *Quantitative Marketing and Economics*, 2011, *9*, 211–257.

Mollick, Ethan, *Co-intelligence: Living and working with AI*, Penguin, 2024.

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

Mozannar, Hussein and David Sontag, "Consistent estimators for learning to defer to an expert," in "International conference on machine learning" PMLR 2020, pp. 7076–7087.

Mullainathan, Sendhil and Ziad Obermeyer, "Diagnosing Physician Error: A Machine Learning Approach to Low-Value Health Care," *The Quarterly Journal of Economics*, May 2022, *137* (2), 679–727.

Olea, José Luis Montiel and Mikkel Plagborg-Møller, "Simultaneous confidence bands: Theory, implementation, and an application to SVARs," *Journal of Applied Econometrics*, 2019, *34* (1), 1–17.

Ostrovsky, Michael and Michael Schwarz, "Reserve prices in internet advertising auctions: A field experiment," *Journal of Political Economy*, 2023, *131* (12), 3352–3376.

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 combinations of humans and AI are useful: A systematic review and meta-analysis," *Nat Hum Behav*, 2024, *8*, 2293–2303.

Vodrahalli, Kailas, Tobias Gerstenberg, and James Y Zou, "Uncalibrated models can improve human-ai collaboration," *Advances in Neural Information Processing Systems*, 2022, *35*, 4004–4016.

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

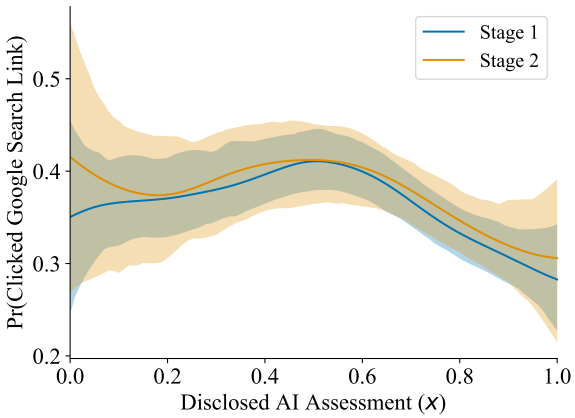
Appendix

A Data Appendix

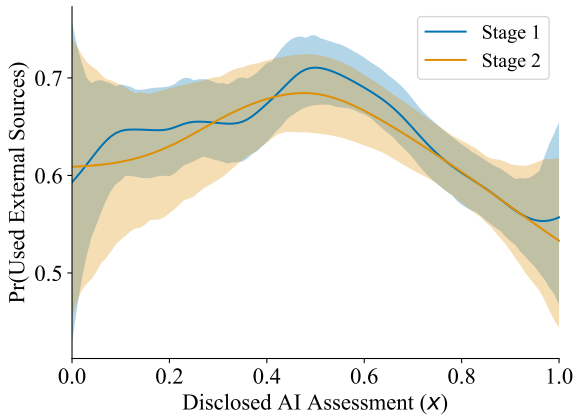
A.1 Effort Response for Additional Effort Measures

Figure A.1: Effort Response for Additional Effort Measures

(a) Clicked Google Search Link



(b) Used External Sources



Note: Plots of additional effort measures conditional on x in the Full Disclosure + No Automation arm. Clicked Google Search Link is an indicator if the participant clicked the provided search link and Used External Sources is an indicator if the participant self-reported using external sources on a case. The curves are estimated via local linear regression and the confidence bands represent bootstrapped 95% uniform confidence bands. The bandwidth is chosen via leave-one-out cross validation.

A.2 Balance Tests

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

Table A.1: Covariate Balance in Stage 2

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

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

A.3 Sample Composition

Table A.2 compares the distribution of study participants in Stage 1 and Stage 2 to the adult U.S. population. Table A.3 provides the funnel of participants throughout the study.

Table A.2: Representativeness of Study Participants

		Stage 1		**Stage 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: Means are computed from Stages 1 and 2. The US Census values are calculated from US Census Bureau population group estimates from 2021 and adjusted to account 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.

Table A.3: Pipeline of Study Participants

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

Note: Table computes the number of participants under various study outcomes. Reached Consent is the number of participants that viewed the consent page. Consented is the number of participants that provided consent. Began Study denotes the number of participants that completed the five practice claims (these participants all completed the instruction pages and comprehension questions). Completed is the number of participants who successfully completed the study without technical issues.

A.4 Robustness

Here we demonstrate the key results are robust to a number of alternative specifications. Table A.4 contains the average accuracy in Stage 2 when focusing on a pure across comparison, including participant and case fixed effects, controlling for the treatment order, and controlling for the number of prior statements checked. The results are quantitatively and qualitatively similar across the various specifications.

Table A.4: Average Accuracy by Treatment

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

Note: This table summarizes estimates of the average treatment effect on accuracy (proportion correct) in Stage 2 for different specifications. Column (1) estimates the treatment effect without controls or fixed effects. Column (2) only uses data from the first treatment encountered for each participant. Column (3) includes participant and case fixed effects. Column (4) controls for treatment order. Column (5) controls for the number of prior claims encountered. Each model is estimated via OLS. In panel B, the outcomes have been adjusted to account for automation as described in footnote 32. Standard errors in parentheses are two-way clustered at the participant and claim level.

A.5 Robustness of Effects on Effort

Here we demonstrate the treatment effect estimates of the various designs on effort (Table 3) are robust to a number of alternative specifications. Table A.5 contains the effects on effort when focusing on a pure across comparison, Table A.6 includes participant and case fixed effects, Table A.7 controls for the treatment order, and Table A.8 controls for the number of prior statements checked. The results are quantitatively and qualitatively similar across nearly all specifications. In the pure across comparison, the constants demonstrate more effort relative to the full sample given participants became faster over time. In addition, the treatment effects are less precisely estimated in the across comparison.

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

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

Note: The average treatment effect is estimated using equation 5. Only the first treatment encountered for each participant is included. This table summarizes the across average treatment effects of different information environments on effort. In panel B, the outcomes have been adjusted to account for automation as described in footnote 32. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

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

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

Note: The average treatment effect is estimated using equation 5 with additional fixed effects at the participant and case levels. This table summarizes the average treatment effects of different information environments on effort. In panel B, the outcomes have been adjusted to account for automation as described in footnote 32. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

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

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

Note: The average treatment effect is estimated using equation 5 plus controlling for treatment order. This table summarizes the average treatment effects of different information environments on effort. In panel B, the outcomes have been adjusted to account for automation as described in footnote 32. Time taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

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

Treatment	**External Sources**	**Clicked Google**	**Time Taken (s)**
	(1)	(2)	(3)
Panel A: No Automation Baseline (β_0)			
Full Disclosure	0.681	0.430	53.473
(FDNA)	(0.009)	(0.010)	(0.835)
Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)			
Full Disclosure	-0.357	-0.209	-24.523
(FDA)	(0.006)	(0.006)	(0.527)
No Disclosure	-0.412	-0.240	-28.497
(NDA)	(0.007)	(0.007)	(0.568)
Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)			
No Disclosure	0.063	0.045	3.690
(NDNA)	(0.006)	(0.006)	(0.534)
Stoplight	0.003	0.001	0.021
(SL)	(0.005)	(0.005)	(0.478)
Observations	80000	80000	80000

Note: The average treatment effect is estimated using equation 5 plus controlling for the number of prior statements assessed. This table summarizes the average treatment effects of different information environments on effort. In panel B, the outcomes have been adjusted to account for automation as described in footnote 32. Time Taken (s) is measured in seconds and winsorized to the 95th percentile. Standard errors are two-way clustered at the participant and claim level in parentheses.

A.5.1 Deviation from Ground Truth

Table A.9: Average Treatment Effects on Accuracy (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
(FDNA)	(0.004)	(0.003)
Panel B: Automation Treatment Effects ($\beta_k - \beta_0$)		
Full Disclosure	0.026	-0.006
(FDA)	(0.004)	(0.003)
No Disclosure	0.024	-0.000
(NDA)	(0.004)	(0.003)
Panel C: No Automation Treatment Effects ($\beta_k - \beta_0$)		
No Disclosure	-0.034	0.032
(NDNA)	(0.005)	(0.003)
Stoplight	0.002	0.001
(SL)	(0.005)	(0.003)
Observations	80000	80000

Note: This table summarizes the treatment effects of different information environments on the assessment accuracy as measured by proportion correct (column (1)) and and deviation from ground truth (column (2)). In panel B, the outcomes have been adjusted to account for automation. Standard errors are two-way clustered at the participant and claim level in parenthesis.

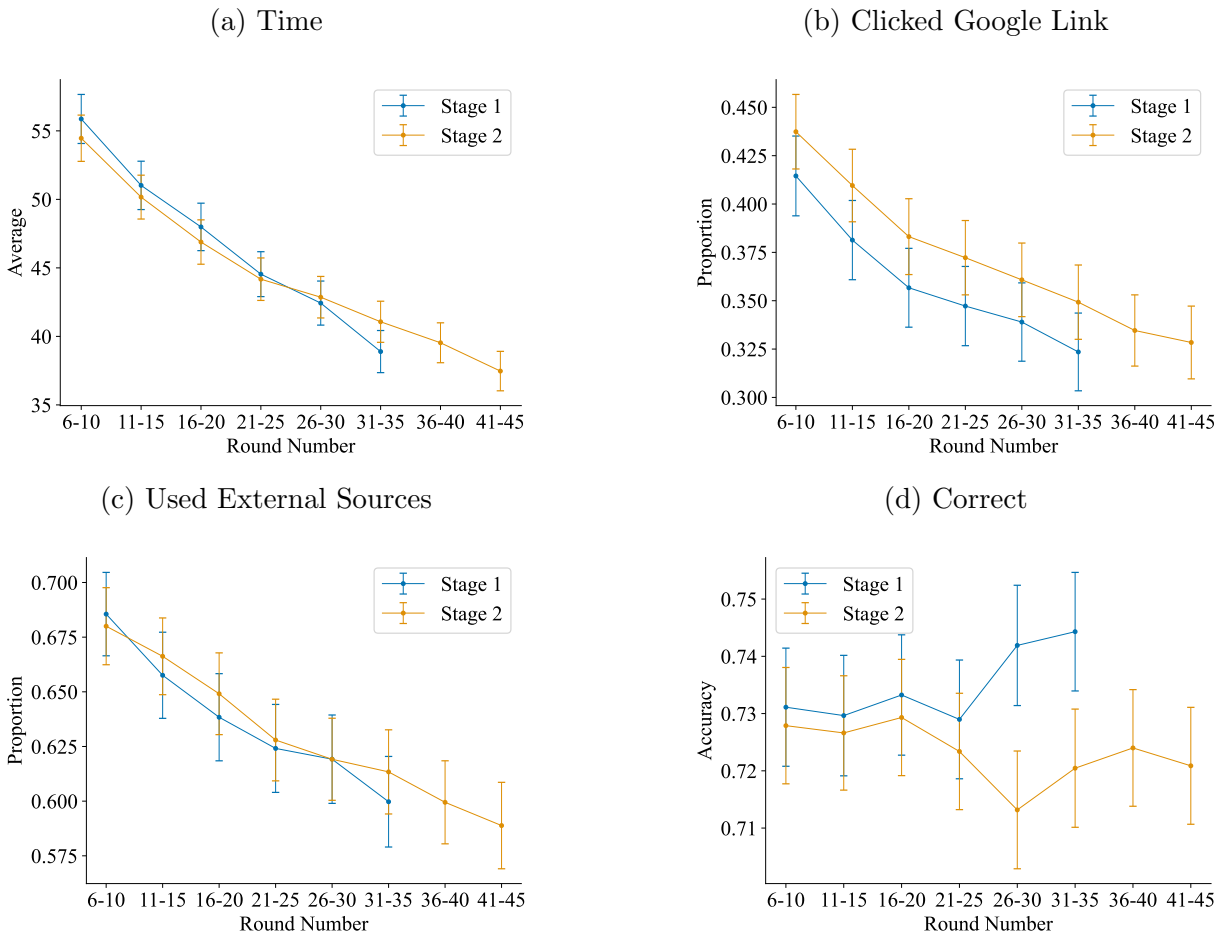
A.6 Fatigue and Learning

Participants classified 30 claims in Stage 1 and 40 claims in Stage 2. In addition, all participants assessed 5 practice claims that we exclude. We test for fatigue and learning effects by estimating the following regression model separately for each stage and plotting β_k in figure A.2

$$y_{i,j} = \sum_{k \in \mathcal{I}} 1[\text{ interval } (i,j) = k]\beta_k + \sum_{g \in \mathcal{G}} 1[\text{ policy } (i,j) = g]\gamma_g + \varepsilon_{ij} \tag{8}$$

where y_{ij} is the outcome for participant i on claim j. The set $\mathcal{I} = \{6\text{–}10, 11\text{–}15, \ldots, 41\text{–}45\}$ splits the claims into eight blocks of five claims, and $\mathcal{G}$ indexes the four policies in Stage 2. FDNA is the omitted treatment, so each γ_g measures the mean difference between treatment g and FDNA. The β_k's represent the learning and fatigue trend after removing these treatment effects.

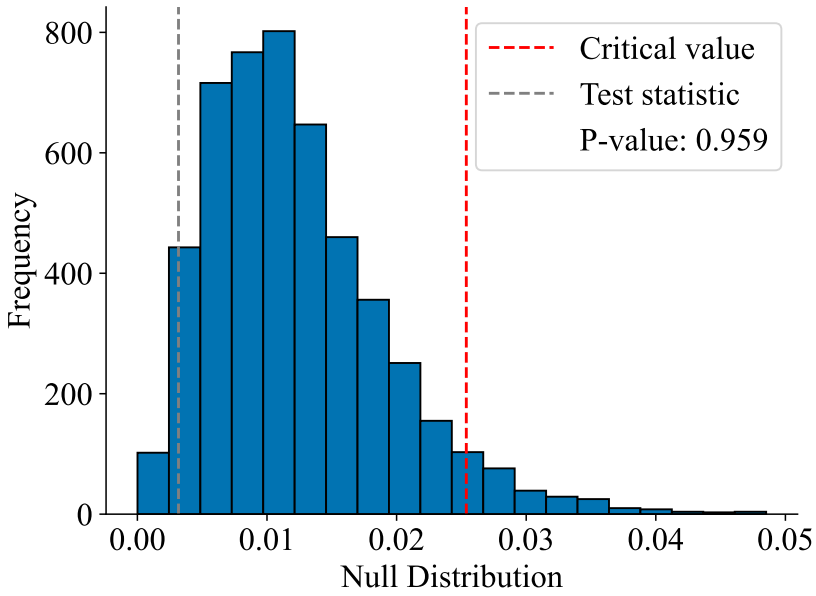
Figure A.2: Outcome by Round Number



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

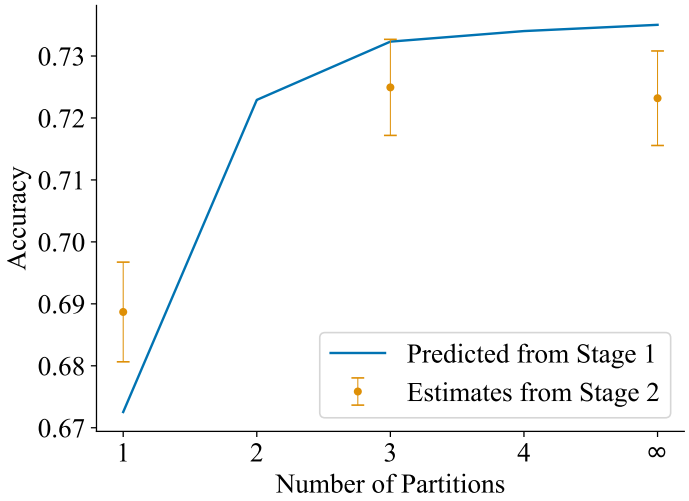
B Balance and Stability of V

Figure B.3: Test for Convexity of $V(\theta)$ versus Null Distribution



Note: We test convexity of $V(\theta)$ estimated from Stage 1 using the bootstrap procedure to test shape restrictions proposed in Fang and Seo (2021).

Figure B.4: Stoplight Policy Predicted Accuracy by K



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

Table B.10: Balance: Stage 1 vs Stage 2

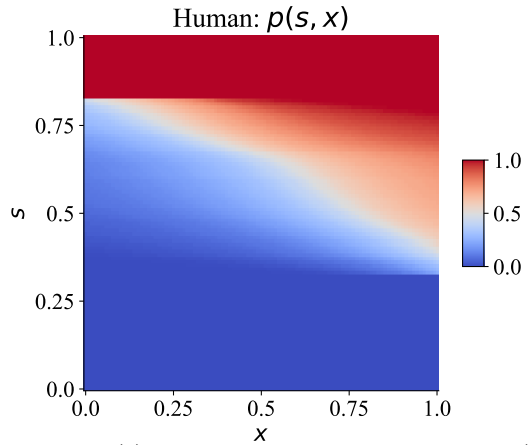
	Stage 1		Stage 2			
	Mean	SD	Mean	SD	Diff	p-value
	(1)	(2)	(3)	(4)	(5)	(6)
Correct Classification	0.735	0.441	0.723	0.447	0.012	0.008
Classified as True	0.696	0.460	0.696	0.460	-0.001	0.912
Assessment	0.630	0.329	0.629	0.318	0.001	0.732
Used External Sources	0.637	0.481	0.630	0.483	0.007	0.579
Clicked 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 Stage 1, while Columns (3) and (4) present the same statistics for Stage 2. Column (5) reports the difference between column (1) and column (3), and column (6) reports the p-value that the difference is statistically significant. The p-value in column (6) is from a regression of the outcome on a constant and Stage 2 indicator, with two-way clustering on participants and cases. Correct Classification is an indicator for whether the decision matches the ground truth. Classified as True is an indicator for whether the probability reported > 0.5. Assessment is the probability true reported. Used External Sources is an indicator for whether the participant self-reported using external sources for a particular case. Clicked Google Link is an indicator for whether the participant clicked on the Google link provided by the experimental interface for a particular case. Time Taken (s) is measured in seconds and winsorized to the 95th percentile.

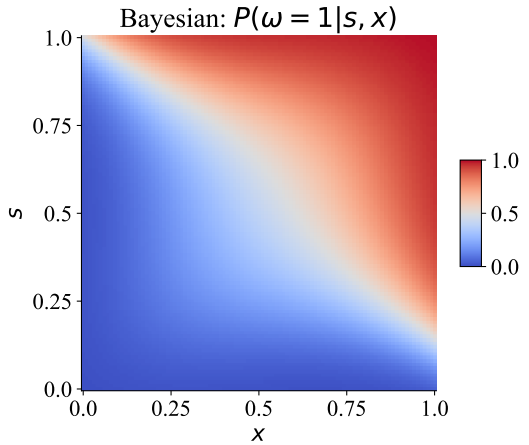
C Human v Bayesian Update Rule

Figure C.5: Human vs Bayesian Update Rule

(a) Humans



(b) Bayesian

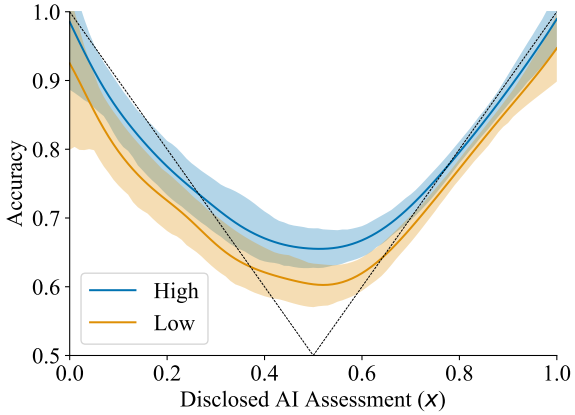


Note: Panel (a) plots the estimate of the function $p(s,x)$ that humans use to combine their own information with the AI assessment. Panel (b) plots the function a Bayesian decision-maker uses to combine the two sources of information $p^{\text{Bayes}}(s,x)$.

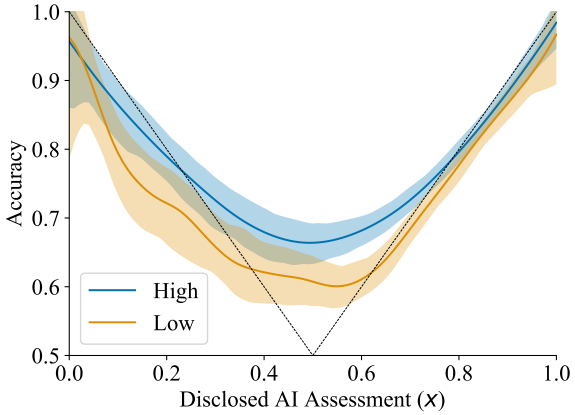
D Heterogeneity in V

Figure D.6: Heterogeneity in $V(\theta)$

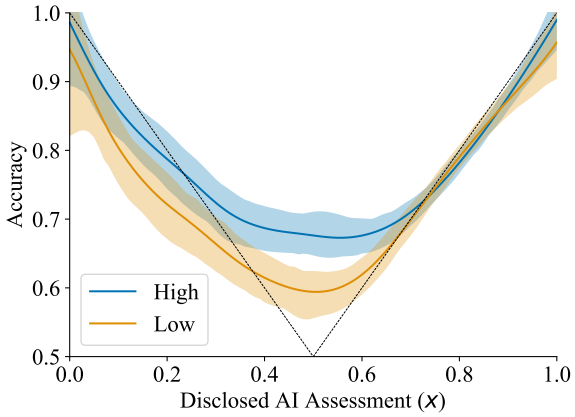
(a) Split by Comprehension



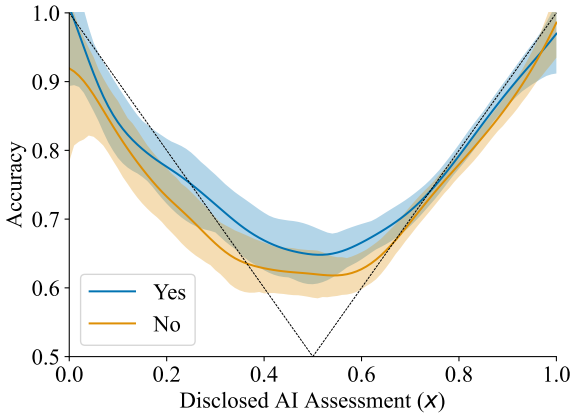
(b) Split by Performance



(c) Split by Effort



(d) Split by Overconfidence



Note: $V(\theta)$ is estimated using local linear regression from Stage 1 data. $V(\theta)$ is estimated separately for high and low conscientiousness participants, and conscientiousness is measured in four ways: (a) number of comprehension questions answered correctly in the training section (two or less wrong indicates high conscientiousness), (b) performance as measured by a regression of correct minus $max\{\theta, 1-\theta\}$ on participant fixed effects, (c) effort as measured by a regression of used external sources indicator on θ, θ^2, and participant fixed effects, and (d) confidence as measured by a regression of the ground truth on a constant, and the probability reported interacted with participant fixed effects. For figures D.6b, D.6c, and D.6d, 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 D.11: Heterogeneity in Predicted Performance

	SL		FDA		NDA	
	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	0.749	0.750	0.759	0.760	0.754	0.754
No	0.722	0.727	0.746	0.746	0.742	0.745

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 Stage 1 data. The "Separate" column denotes the performance of individually estimated policies for each group (high and low by comprehension, performance, effort, and overconfidence, resulting in 8 unique $V(\theta)$ curves) using the unique $V(\theta)$.

E Alternative Designer Preferences

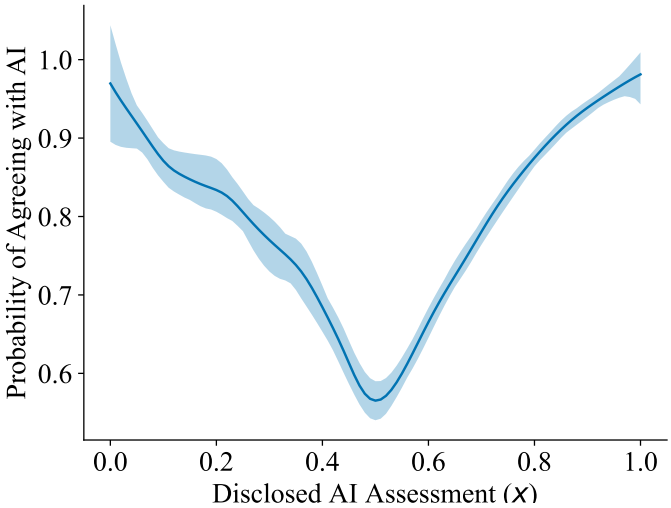
We now estimate V for alternative designer preferences. First, we consider a designer who wants to persuade human decision-makers to follow the AI recommendation. This situation could arise if the human decision-maker's signal is a garbling of the AI signal, as in Dreyfuss and Hoong (2025). In this case, the Bayesian posterior is equal to the AI's assessment, which may differ from humans' assessments if they deviate from Bayesian updating.

Figure E.7 shows the estimate of the probability that a human decision-maker follows the AI assessment conditional on x. This function is not convex, implying that a designer who maximizes the probability that the decision-maker agrees with the AI benefits from coarsening the AI assessment. In particular, the optimal policy without automation for this indirect utility function is a binary signal that pools all AI assessments between 0 and 0.52 into one signal, and pools all AI assessments between 0.53 and 1 into a second signal. This is consistent with the empirical results in Dreyfuss and Hoong (2025), which finds that a binary coarsening improves performance over full disclosure in a setting where the human's information is a subset of the AI's.

Next, we consider a designer who wants to minimize the expected deviation of the decision-maker's probability assessment from the ground truth. Figure E.8 shows the estimate of this indirect utility function conditional on x. This function is convex, implying that a designer

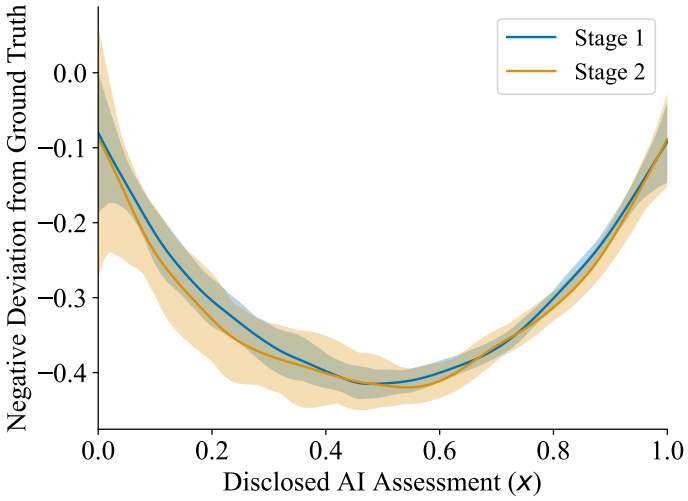
with this objective should fully disclose the AI assessment.

Figure E.7: V Defined Using Probability of Agreeing with AI



Note: Here $V(x)$ is defined as $\Pr(a = 1[\theta > 0.5]|x)$. V is estimated using local linear regression separately using Stage 1 data. The 95% uniform confidence band displayed is computed via bootstrap accounting for clustering at the participant and case level.

Figure E.8: V Defined Using Deviation from Ground Truth



Note: Here $V(x)$ is defined as $-E\left[|p_{ij} - \omega_i| \ |x\right]$. V is estimated using local linear regression separately using Stage 1 data and Stage 2 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.

F Alternative Design Approaches

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

F.1 Algorithmic Triage Approach

The algorithmic triage literature focuses on algorithms that selectively automate cases and assign the remaining cases to human-decision-makers without considering how humans' decision accuracy responds to the automation policy. The sufficient statistic approach has two main differences from the algorithmic triage approach. First, as discussed in the text, the sufficient statistic approach has a lighter data requirement for calculating the optimal automation/disclosure policy. Second, the sufficient statistic approach accounts for changes in humans' beliefs in response to the designer's policy. This leads to quantitatively different predicted 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 without disclosing the AI asssessment. The optimal one-sided automation policy automates cases where $\theta > 0.58$. We can calculate the predicted accuracy of this policy as $\gamma^H Pr\left(\theta \leq 0.58\right) + \mathbb{E}\left[\theta | \theta > 0.58\right] Pr\left(\theta > 0.58\right)$, where γ^H is the predicted accuracy of humans on cases assigned to them. The sufficient statistic approach predicts $\gamma^H = V\left(\mathbb{E}\left[\omega | \theta \leq 0.58\right]\right) = 65.3\%$, while the algorithmic triage approach treats human performance as fixed and predicts $\gamma^H = E\left[1\left[\omega_i = a_{ij}\right] | \theta \leq 0.58\right] = 61.2\%$ using data from the NDNA arm. The difference in performance results because the sufficient statistic approach assumes that humans' beliefs depend on the distribution of cases they encounter in response to the automation policy.

F.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. However, 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 naïve 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

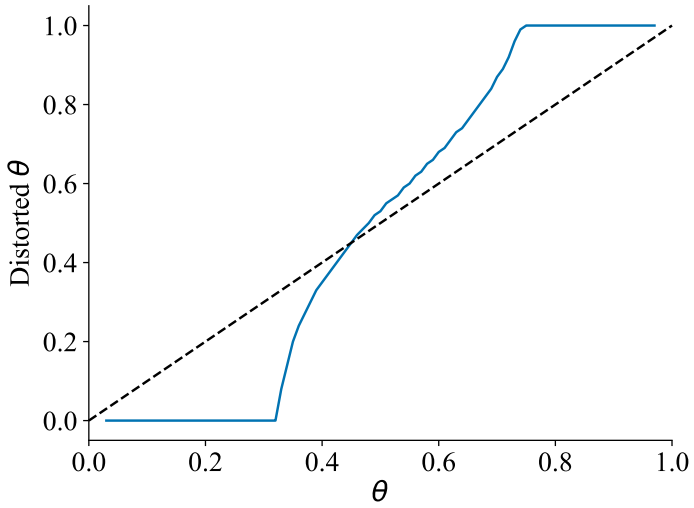
$$\theta T(\delta(\theta), 1) + (1 - \theta)(1 - T(\delta(\theta), 0)), \tag{9}$$

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

It is straightforward to solve the naïve designer's problem and compare its naïve and sophisticated expected accuracy. We estimate the function $T(\theta, \omega)$ using a logistic regression with a quadratic term on θ and solve the optimal distortion problem of the naïve designer. Figure F.9 plots the naïve optimal distortion policy $\delta(\theta)$. Due to the AI under-response we have documented throughout the paper, the naïve designer exaggerates the AI signal, for example by reporting $\delta(\theta) = 1$ whenever $\theta \geq 0.75$ and reporting $\delta(\theta) = 0$ whenever $\theta \leq 0.32$.

This naïve optimal policy yields a naïve expected accuracy of 74.7%. This accuracy is very close to that under Full Disclosure + Automation (75.1%). Intuitively, a naïve designer believes that she can nearly replicate automation by exaggerating signals where the AI is confident. However, the sophisticated expected accuracy of this policy is only 73.3%, which is worse than the expected accuracy of 73.5% under FDNA. 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.

Figure F.9: Naïve Designer Distortion Map



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

G Estimating Conditional Probabilities

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

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

where β is a vector of parameters, $p(W_{ij}, \beta) = \frac{\exp(\beta' W_{ij})}{1+\exp(\beta' W_{ij})}$, $||\cdot||_1$ represents the l_1 norm, $||\cdot||_2$ represents the l_2 norm, and λ_1, λ_2 are non-negative tunable parameters.

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

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

H Analyzing Endline Responses

We asked 5 endline questions to participants and summarized these responses using Claude Sonnet 4. This analysis was not pre-registered and is intended to provide additional insight. Specifically, we prompted the model with, "I ran a study to design human-AI collaboration for a fact-checking task. Please summarize the debrief questions and key themes. Include representative quotes when possible."

Only 3,498 of the 3,501 participants completed the endline questions. The remaining text in this section contains the un-edited text generated by the large language model.

Debrief Questions

The study collected responses from 3,498 participants across five key debrief questions:

1. **Fact-Checking Strategy**: "How did you try to determine the likelihood a statement was True or False? Were there any clues or resources that were particularly useful?"
2. **External Resource Usage**: "What outside resources did you use? What was your strategy when using them?"
3. **AI Collaboration Approach**: "How did you use the AI to solve the task? What was your strategy?"
4. **Study Purpose Perception**: "What did you think is the purpose of the study?"
5. **General Feedback**: "In the space below, please feel free to share any thoughts you have on the study."

Five Key Themes

Heavy Reliance on Google Search and Wikipedia

Participants overwhelmingly turned to Google searches and Wikipedia as their primary fact-checking resources, often viewing these as essential rather than supplementary tools.

Representative Quotes:

- "I couldn't find a way to determine the likelihood without using a google search."
- "Wikipedia was an especially helpful resource, although at times what Wikipedia said was different from what the correct answer was."
- "Wikipedia everytime, looked at dates"

- "I used wikipedia it seemed to help."

Varied AI Trust and Utilization Strategies

Participants showed a spectrum of approaches to using AI assistance, from complete reliance to active skepticism, with many developing threshold-based decision rules.

Representative Quotes:

- "I trusted AI if there was above 50 probability for the model being correct."
- "I consider the AI's likelihood percentage as a primary clue, which indicates the average probability based on its evaluation."
- "I stopped looking at the AI's statement likelihood"
- "I looked at the AI evaluator's percentage at times when I was stuck to see what its evaluation was and just went with that answer."

Keyword-Based Search Strategies

Participants developed systematic approaches to fact-checking, focusing on identifying key terms, dates, and specific phrases to guide their searches.

Representative Quotes:

- "I used a combination of the AI's evaluation and several specific google searches containing key words from each phrase."
- "I would put certain phrases in quotes in google search so that it would include resources containing those phrases."
- "Using the external source and including key terms (a year, for example)."
- "I tried to determine whether it was true or false by doing my search and looking for certain keywords to use as clues."

Clear Understanding of Study Purpose

Most participants accurately identified the study's focus on AI accuracy evaluation and human-AI collaboration in fact-checking tasks.

Representative Quotes:

- "To see how effectively people fact check phrases that AI deems to be correct or false."

- "To see how much people feel they can trust AI."
- "how accurate AI is."
- "To see how we research?"

Mixed Attitudes Toward AI Performance

Participants expressed a range of sentiments about AI capabilities, from enthusiasm to skepticism, with many providing constructive feedback about the collaboration experience.

Representative Quotes:

- "AI is dogwater" (expressing frustration)
- "It was engaging for sure!" (positive engagement)
- "Thank you." (appreciation)
- "At times, though, I needed to verify that the AI evaluator was correct with a google search." (cautious verification approach)

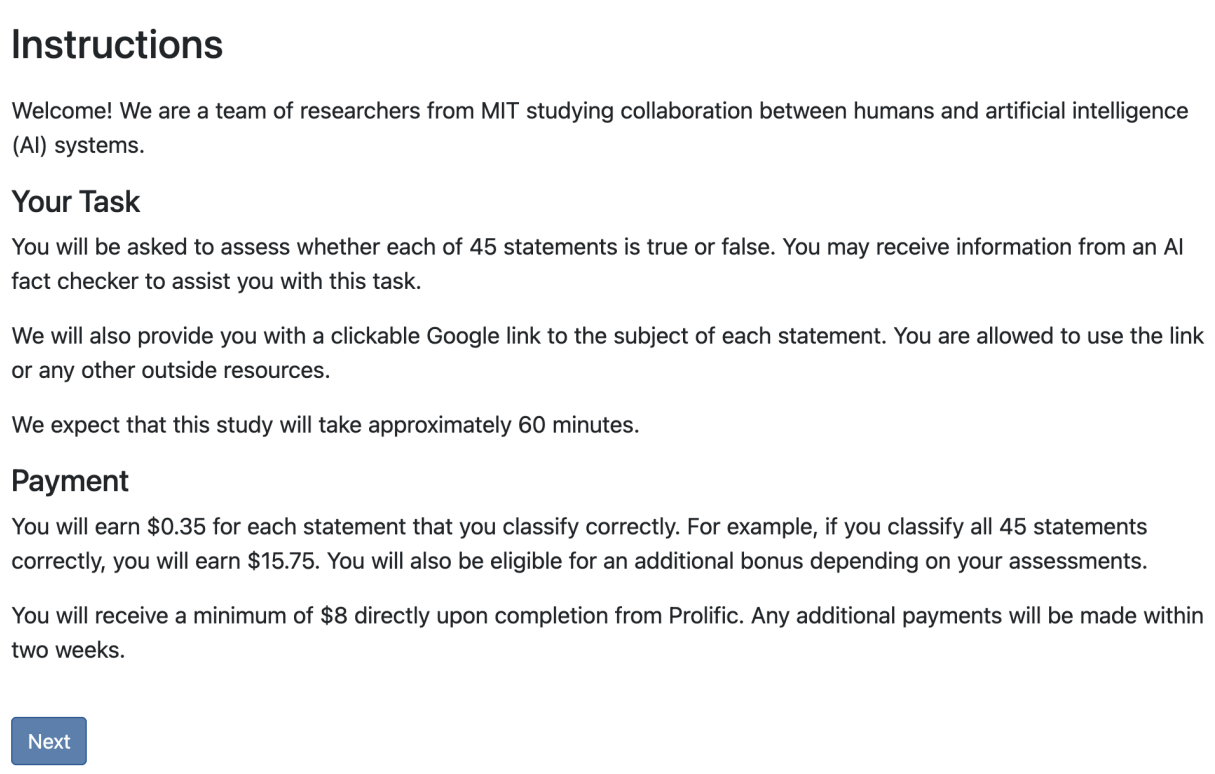
Summary

The study reveals that participants developed sophisticated fact-checking strategies combining AI assistance with traditional web search methods. While Wikipedia and Google emerged as trusted verification sources, participants showed varying levels of AI trust, often developing personal thresholds for when to rely on AI predictions versus conducting independent verification. The collaborative approach highlighted both the potential and limitations of human-AI partnerships in fact-checking tasks.

I Experimental Instructions

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

I.1 Instruction Page 1



I.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

I.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

I.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

I.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

I.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