A Sufficient Statistic for Designing and Evaluating Human-AI Collaboration (Pre-Analysis Plan)

Nikhil Agarwal, Alex Moehring, Alexander Wolitzky

July 10, 2024

1 Introduction

Predictive AI tools have recently matched or surpassed human performance across numerous domains [Liu et al., 2019, Lai et al., 2021, Mullainathan and Obermeyer, 2019, Kleinberg et al., 2017, Agrawal et al., 2018]. Yet, in a number of settings, there may be value from collaboration between humans and AI tools, particularly if humans have access to alternative sources of information [Agarwal et al., 2023]. This possibility motivates the design of collaboration systems that can improve performance over a "human-only" or an "AI-only" system by fully taking advantage of all available information.

Such a design should be based not only on the relative performance of human and AI predictions, but also on how potential human biases and incentives to exert effort in response to AI assistance affect performance. Existing empirical approaches to designing human-AI collaboration typically abstract away from these endogenous responses.[1] One approach for accounting for these changes is to run experiments under various designs, which is prohibitively expensive given the large number of potential designs. Another approach would be to build a structural model of behavior with built-in assumptions about how humans respond to the design.

We take a third approach based on a new sufficient statistic that can be used to design human-AI collaboration. Our theoretical framework is outlined in section 2. It characterizes the optimal design of human-AI collaboration when it is possible to automate certain cases or to partially withhold AI-generated information. The experiment demonstrates how to apply the framework to design a collaborative system.

Our experiment will address the following research questions:

1. What is the optimal use of AI predictions in collaboration with human decision-makers?

2. Does fully revealing the AI prediction to human decision-makers maximize performance?

3. Which types of decisions should be automated?

[1] A literature in computer science considers whether or not to automate certain tasks to humans [Mozannar and Sontag, 2020, Raghu et al., 2019, Bansal et al., 2021] based on relative performance alone. Within economics, a recent set of papers papers compare humans to AI or humans with AI to AI alone [Liu et al., 2019, Lai et al., 2021, Mullainathan and Obermeyer, 2019, Kleinberg et al., 2017, Agrawal et al., 2018], and the effect of biases in humans' use of AI predictions on human-AI collaboration [Agarwal et al., 2023]. However, there is little work on how AI predictions shape humans' incentives to gather their own information and the implication on human AI collaboration. An exception is Athey et al. [2020], which presents a theoretical reason why AI assistance is not always optimal.

4. Do humans respond to confident AI predictions by reducing effort in gathering their own information? If so, is it optimal to withhold information in AI predictions in order to incentivize effort?

5. Is the sufficient statistic approach valid, i.e. stable under alternative collaborative designs?

The experimental design requires the data collection to proceed in two phases. A first phase will be used to estimate primitive objects that can be used to calculate the optimal design and predict the performance of alternatives. The second phase will experimentally test a set of optimal and constrained designs, and compare them with a benchmark. Further details of the experimental procedures are presented in section 3.

We study these questions in the context of fact-checking. Fact-checking provides an ideal setting to investigate how to design human-AI collaboration. The veracity of information people consume online has become increasingly important to policy-makers and researchers around the world [Lazer et al., 2018]. One approach to limit the spread of false information online relies on fact-checkers. This approach is taken by many large digital platforms including Facebook [Facebook]. Recently, researchers across disciplines have focused on improving the productivity of human fact-checking systems [Allen et al., 2021] and automating parts of the process [Guo et al., 2022].

In addition to being an important setting for understanding how to design the optimal provision of AI assistance to humans, fact-checking is also convenient for experimental purposes. Measuring performance in fact-checking process is relatively simple, with a clearly defined binary state of the world of a statement either being true or false. There are also established datasets and benchmarks containing both datasets of true and false claims (e.g. Aly et al. [2021]) and manually curated ground-truth labels. In addition, recent work has suggested crowd-workers can be effective fact-checkers [Allen et al., 2021]. Moreover, AI tools for fact-checking can be readily developed.

2 Theoretical Framework

The experiment estimates the parameters needed to design an optimal policy that combines information disclosure from an AI system to a human decision maker and automation of decisions by the AI system in order to maximize the expected performance of the overall human-AI collaborative system. This section describes the theoretical framework underlying this problem and the parameters to be estimated.

We study binary classification problems: there is a binary state of the world $\omega \in \{0,1\}$ (whether a given statement is false or true) and a binary classification decision $a \in \{0,1\}$ (whether the statement is classified as false or true). For each statement, the AI determines an assessment $\theta \in [0,1]$ of the probability that $\omega = 1$. The assessment is calibrated: $\Pr(\omega = 1|\theta) = \theta$. Denote the population distribution of the AI assessment θ by F. For each statement, the AI then either discloses a signal of its assessment to the human subject or makes the classification on its own. We assume that the probability that a human subject makes the correct classification when they learn that the mean AI assessment is x is well-defined, and we denote this probability by $V(x)$. For example, $V(x)$ is well-defined if subjects are Bayesian and their own information about ω is conditionally independent of θ. Let $W(x) = \max\{V(x), 1-x, x\}$, which is the maximum performance attainable by an AI with assessment x by either disclosing this assessment to the subject ($V(x)$), classifying the statement as false without human input ($1-x$), or classifying the statement as true without human input (x). It can be shown that the maximum expected performance attainable by any AI system in this setting is

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

where $MPC(F)$ denotes the set of all distributions that are *mean-preserving contractions* of the distribution of AI assessments F. The optimal policy is then given by (i) coarsening the AI's assessment so that the distribution of coarsened assessments x is given by the solution G, (ii) disclosing the coarsened assessment x if $V(x) \geq \max\{1-x, x\}$, and (iii) classifying the statement as false (resp., true) without human input if $x < \min\{1-V(x), 0.5\}$ (resp., $x > \max\{V(x), 0.5\}$).

Similarly, the maximum expected performance attainable by information disclosure alone (when the AI is not permitted to make the classification on its own) is

$$V^* = \max_{H \in MPC(F)} \int_0^1 V(x)\, dH(x)\,.$$

The parameters of the framework are thus the distribution of calibrated AI assessments F and the function $V(x)$ describing the performance of human participants as a function of the disclosed mean AI assessment x. In our experiment, the distribution of assessments F is given and known. The experiment thus estimates the function $V(x)$. Given this function, we can calculate the optimal mixed disclosure/automation policy G and the optimal disclosure-only policy H as described above. We can also compare the values of these policies with that of the *full disclosure* policy, where the AI always discloses its assessment, which is given by $\int_0^1 V(x)\, dF(x)$, and that of the *no disclosure* policy, where the AI reveals no information, which is given by $V\left(\int_0^1 x dF(x)\right)$.

3 Experimental Design

3.1 Sample of Claims

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

FEVEROUS conducted extensive quality control to ensure the creation of high quality claims and labels. We refer readers to Aly et al. [2021] for full details of this process. In addition to the quality control measures taken in Aly et al. [2021] we remove claims that are not suitable for our study. Table 1 displays how many claims are removed in each cleaning step. We first remove approximately 3% of claims with a ground truth label of Not Enough Information. We then remove claims with any spelling or grammatical errors flagged by either the rules-based LanguageTool API or GPT-4o.[2][3] Finally, we remove claims that we determine to be poor quality, which primarily consists of claims where the ground truth can change over time, such as for claims that reference an individual's age.

3.2 AI Fact Checker

We use OpenAI's GPT-4o as an automated fact-checker. Specifically, for each of the 41969 final claims used in the analysis we query the OpenAI API with the prompt "True or False: [claim]" and store the top 20

[2] We use the language_tool_python package (`https://github.com/jxmorris12/language_tool_python`) that is a wrapper for the LanguageTool API (`https://languagetool.org`).

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

Table 1: Claim-Cleaning Funnel

	N	Share
Total	78982	1.00
Filter Not Enough Info	76248	0.97
Filter Spelling / Grammar	42707	0.54
Filter Bad Facts	41969	0.53
Final	41969	0.53

most likely next tokens along with the probability distribution GPT-4o assess over these possible tokens. We calculate a score a_i for each claim i as follows

$$a_i = \frac{\sum_j p_j 1\left[\text{token}_j = \text{true}\right]}{\sum_j p_j 1\left[\text{token}_j \in \{\text{true}, \text{false}\}\right]}$$

where token_j is the canonical form (i.e. lower case) of the j^{th} most likely next token and p_j is the probability GPT-4o assigns over the j^{th} token.

While GPT-4o gives us a set of scores a_i for each claim, the model is not calibrated [Achiam et al., 2023]. Therefore, we calibrate the model to our data by binning the claims by a_i into 200 bins and calculating the share of claims that are true in each bin. This gives us a calibrated AI signal θ_i for each claim i. Figure 1 summarizes the signals generated by the AI fact-checker.

3.3 Experiment Structure

This experiment will occur in two rounds. The first round estimates $V(x)$, the primitive that will be used to calculate the optimal policies and to simulate the performance of alternative policies. The second round tests the performance of participants under several disclosure and automation policies relative to the full disclosure policy and their predicted performance. In each round, we will recruit participants from the Prolific survey platform and ask them to assess the likelihood that each of a random subset of statements is true. Participants will assess statements in different information environments further described below. In each information environment participants may receive assistance from an AI fact-checker. In every round, participants will first assess 5 practice cases under full information to familiarize themselves with the task and interface. These cases will be discarded in all primary analyses.

3.3.1 First Round

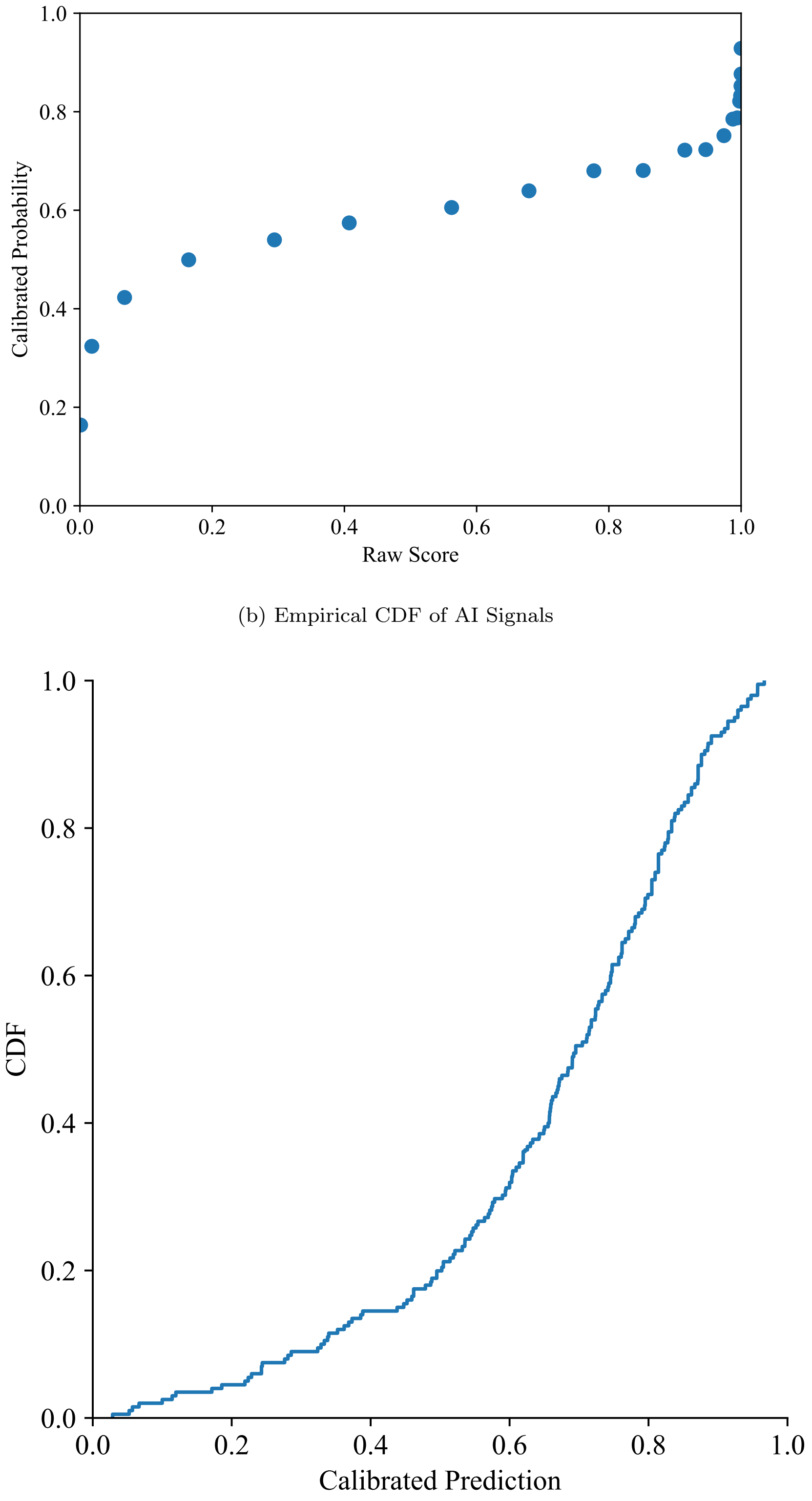
We will estimate $V(\theta)$ using data from 1500 participants that will each assess 30 statements under the full disclosure policy where they observe the AI assessment θ directly. This estimate will allow us to calculate the optimal disclosure policy. Power calculations based on pilot data suggest that the maximum standard error for $V(\theta)$ is at most two percentage points for θ discretized into twenty-one equally spaced bins.

3.3.2 Second Round

In the second round, we will ask 2000 participants to assess claims under varying information environments. This design will include both a within and across comparison. This will allow us to take advantage of the

Figure 1: Summary of AI Predictions

(a) Calibration Function



(b) Empirical CDF of AI Signals

within comparisons to control for heterogeneity in participant skill while preserving the ability to do pure across participant comparisons. Specifically, each participant will assess 10 claims in each of the following information environments

1. No disclosure: participants observe only the prior mean $E[\omega]$
2. Full disclosure: participants observe the AI assessment θ
3. Automation + no disclosure: cases where the AI performance exceeds the human performance are automated and humans observe only the share of cases that are true among those that are not automated.
4. Optimal automation and disclosure: participants observe the signal from the optimal policy G described in Section 2.

The order of the treatments will be randomly drawn for each participant. If the optimal disclosure policy without automation looks sufficiently different from the full disclosure policy (i.e. H from Section 2) we may also include this policy in round 2, and may omit one of the policies listed above. We will amend this pre-analysis plan describing the policies we test in the second round in greater detail once the optimal policies are calculated. Our power calculations suggest that this sample size will be powered at 80% to detect a difference between policies of 1.4 percentage points with a p-value of 5%.

3.4 Variables

For every case we will collect the the participant's assessment $s \in [0,1]$ of the probability a claim is true, whether they click a link to google the subject of the claim, a self-reported measure of whether they used any external sources, and the time taken on the claim (winsorized at the 5th and 95th percentiles). We set the participant's classification of the case to be true if they report $s > 0.5$, i.e. $a = 1\{s > 0.5\}$. Our primary outcomes will be an indicator for whether the participant correctly classified the statement as true or false, and the deviation of a participant's assessment of the likelihood a statement is true from the ground truth. Figure 2 contains a screenshot of the interface used to collect the primary outcome s.

4 Empirical Strategy

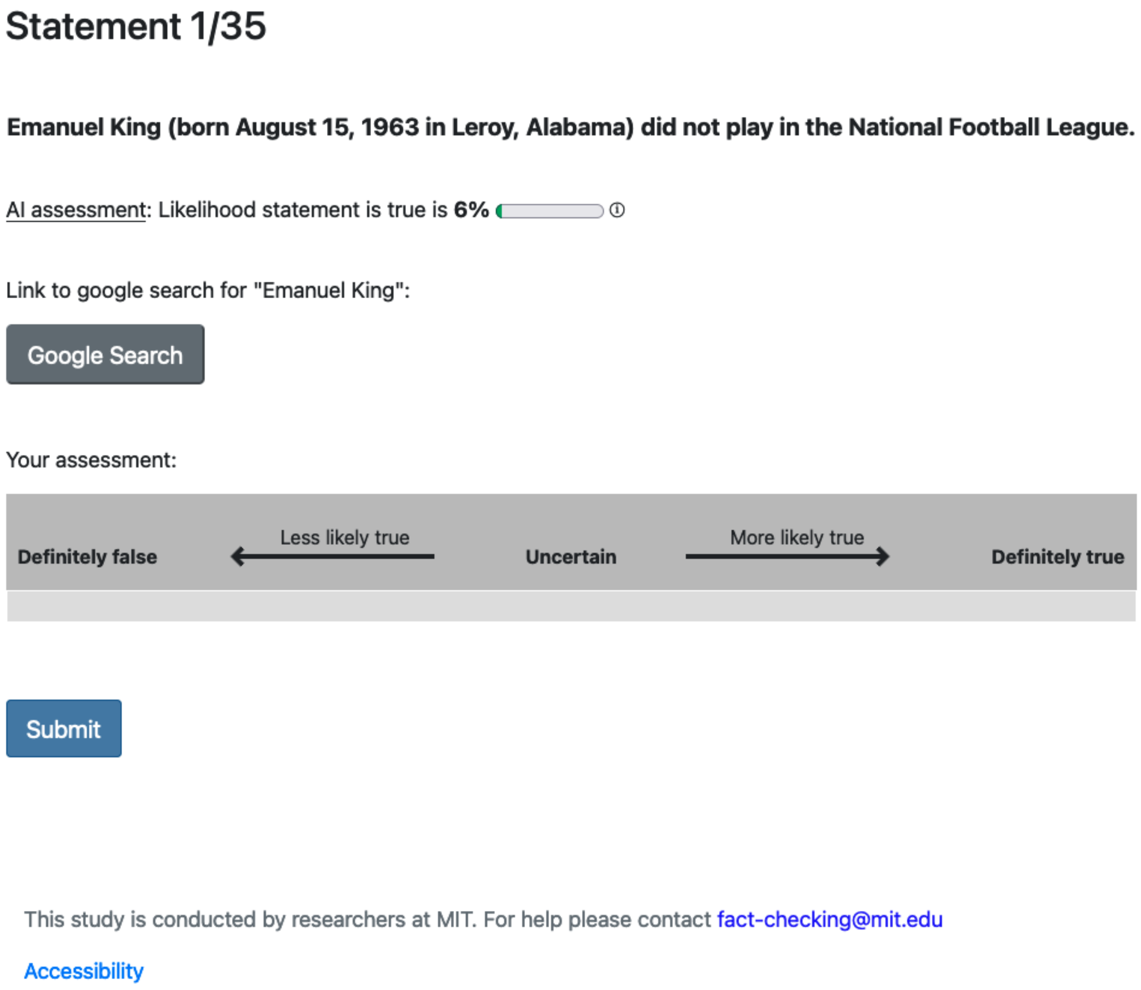
4.1 Round 1

We will estimate $V(\theta)$ using non-parametric regression of accuracy on θ, where accuracy is measured both in terms of the share of statements correctly classified $(1\{a = \omega\})$ and the deviation of the participant's assessment from the ground truth $(|s - \omega|)$. These estimates will then be used to calculate the optimal disclosure and automation policies described in Section 2.

The shape of the function $V(\theta)$ has important implications for the optimal policy. In particular, if $V(\theta)$ is convex it is always optimal to fully disclose θ on cases that are not being automated. Therefore, we will test whether or not $V(\theta)$ is convex.

We will also report non-parametric estimates of effort as a function of the AI signal θ. That is, we will non-parametrically estimate $E[y|\theta]$ where y is an indicator of whether the participant used external sources, an indicator of whether the participant clicked the link to google the statement's subject, and the time the participant took to make their assessment.

Figure 2: Screenshot of Experimental Interface



Statement 1/35

Emanuel King (born August 15, 1963 in Leroy, Alabama) did not play in the National Football League.

AI assessment: Likelihood statement is true is **6%**

Link to google search for "Emanuel King":

Your assessment:

This study is conducted by researchers at MIT. For help please contact fact-checking@mit.edu

Accessibility

4.2 Round 2

In the second round we will estimate the treatment effect of the various disclosure policies on the accuracy of participant's assessments by estimating regressions of the form

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

where y_{ij} is the outcome for statement i and participant j, β_0 is a constant that represents average performance in the full disclosure policy, *Policies* represents the set of policies tested (excluding the full disclosure policy), $policy\left(i, j\right) \in Policies$ indicates the policy under which participant j assessed statement i, β_k represents the average treatment effect of policy k relative to full disclosure, and ε_{ij} is an error term. We will test whether each β_k differ from 0 in addition to testing whether the β_k are statistically distinguishable from one another. All statistical inference will be clustered at the participant and statement level.

The primary outcomes for this analysis will be measures of accuracy including both an indicator of whether the participant correctly classified a statement $y_{ij} = 1\left\{a_{ij} = \omega_i\right\}$ and the deviation of the participant's continuous assessment from the ground truth $y_{ij} = \left|s_{ij} - \omega_i\right|$. Secondary outcomes will again be measures of participant effort including an indicator of whether the participant used external sources, an indicator of whether the participant clicked the link to google the statement's subject, and the time the participant took to make their assessment.

In addition, we will test the stability of $V\left(\theta\right)$ by comparing the performance of the policies in round 2 to the predicted performance under the assumption that $V\left(\theta\right)$ is stable. We will test both the average performance across policies and the function $V\left(\theta\right)$ itself on the support of G. For example given the optimal policy G_1 obtained from the first round, we will test if

$$E\left[y_{ij} | policy\left(i, j\right) = optimal\right] = \int_0^1 W_2\left(x\right) dG_1\left(x\right)$$

where the left hand is the average performance of participants in the optimal policy condition and $W_2\left(\cdot\right)$ is an estimate of $W\left(\cdot\right)$ in the full disclosure condition in the second round. We will also test the null hypothesis that predicted performance equals actual performance at each point in the support of the optimal policy:

$$E\left[y_{ij} | policy\left(i, j\right) = optimal, \theta_i = x\right] = W_2\left(x\right) \;\; \forall x \in Support\left(G_1\right).$$

We will run analogous tests for the other policies tested.

4.3 Additional Secondary Analyses

4.3.1 Simulating the Performance of Alternative Disclosure Policies

Given the estimate of $V\left(x\right)$ from the first round we can simulate the performance of many alternative policies. For example, we will simulate the performance of a policy that discloses at most three AI signals to assess how close the performance of a rule-of-thumb High/Medium/Low policy is to that of the optimal policy.

4.3.2 Estimating the Cost of Human Effort Response

A benefit of collecting continuous assessments is that we can disentangle the relative costs of human shirking and incorrect belief updating. In particular, we can estimate a Bayesian benchmark that optimally combines the assessments of humans in the no disclosure condition with the AI signal. This benchmark holds efforts fixed because the AI signal is not disclosed. Then, we calculate the cost of the human effort responses to AI assistance under correct belief updating by comparing this benchmark to a Bayesian with an effort response. To construct this alternative, we will estimate the optimal combination between our participants' reported assessments in the full disclosure condition with the AI signal. Finally, we can compare these two benchmarks to the assessments made by participants in the full disclosure condition to evaluate the cost of incorrect belief updating.

4.3.3 Estimating the Value of AI Improvement

We will also simulate the performance of different policies under alternative AI's. Specifically, we will study how the optimal policies change if we use the distribution of calibrated AI signals generated from less capable AI fact-checkers.

References

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

N. Agarwal, A. Moehring, P. Rajpurkar, and T. Salz. Combining human expertise with artificial intelligence: Experimental evidence from radiology. Technical report, National Bureau of Economic Research, 2023.

A. Agrawal, J. Gans, and A. Goldfarb. *Prediction Machines: The Simple Economics of Artificial Intelligence*. Harvard Business Press, Apr. 2018.

J. Allen, A. A. Arechar, G. Pennycook, and D. G. Rand. Scaling up fact-checking using the wisdom of crowds. *Science advances*, 7(36):eabf4393, 2021.

R. Aly, Z. Guo, M. S. Schlichtkrull, J. Thorne, A. Vlachos, C. Christodoulopoulos, O. Cocarascu, and A. Mittal. FEVEROUS: Fact extraction and VERification over unstructured and structured information. 2021.

S. C. Athey, K. A. Bryan, and J. S. Gans. The allocation of decision authority to human and artificial intelligence. In *AEA Papers and Proceedings*, volume 110, pages 80–84. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2020.

G. Bansal, B. Nushi, E. Kamar, E. Horvitz, and D. S. Weld. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 11405–11414, 2021.

Facebook. Third-party fact-checking program by facebook. `https://www.facebook.com/formedia/mjp/programs/third-party-fact-checking`.

Z. Guo, M. Schlichtkrull, and A. Vlachos. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206, 2022.

J. Kleinberg, H. Lakkaraju, J. Leskovec, J. Ludwig, and S. Mullainathan. Human decisions and machine predictions. *Q. J. Econ.*, 133(1):237–293, Aug. 2017.

V. Lai, C. Chen, Q. Vera Liao, A. Smith-Renner, and C. Tan. Towards a science of Human-AI decision making: A survey of empirical studies. Dec. 2021.

D. M. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, et al. The science of fake news. *Science*, 359(6380):1094–1096, 2018.

X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern, J. R. Ledsam, M. K. Schmid, K. Balaskas, E. J. Topol, L. M. Bachmann, P. A. Keane, and A. K. Denniston. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The Lancet Digital Health*, 1(6): e271–e297, Oct. 2019.

H. Mozannar and D. Sontag. Consistent estimators for learning to defer to an expert. In *International conference on machine learning*, pages 7076–7087. PMLR, 2020.

S. Mullainathan and Z. Obermeyer. A machine learning approach to low-value health care: wasted tests, missed heart attacks and mis-predictions. 2019.

M. Raghu, K. Blumer, G. Corrado, J. Kleinberg, Z. Obermeyer, and S. Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220*, 2019.