Blueprint Labs

Discussion Paper #2023.10

Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology

Nikhil Agarwal
Alex Moehring
Pranav Rajpurkar
Tobias Salz

July 2023

MIT Department of Economics
77 Massachusetts Avenue, Bldg. E53-390
Cambridge, MA 02139

National Bureau of Economic Research
1050 Massachusetts Avenue, 3rd Floor
Cambridge, MA 02138

NBER WORKING PAPER SERIES

COMBINING HUMAN EXPERTISE WITH ARTIFICIAL INTELLIGENCE: EXPERIMENTAL EVIDENCE FROM RADIOLOGY

Nikhil Agarwal
Alex Moehring
Pranav Rajpurkar
Tobias Salz

Working Paper 31422
http://www.nber.org/papers/w31422

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
July 2023

The project benefitted from collaboration with several radiologists, including Drs. Matthew Lungren, Curtis Langlotz, and Anuj Pareek of Stanford, Drs. Etan Dayan and Adam Jacobi of Mt. Sinai Hospital, Steven Truong of VinBrain and several radiologists at VINMEC, and teleradiologists at USARAD, Vesta Teleradiology, and Advanced Telemed. We thank Daron Acemoglu, David Autor, David Chan, Glenn Ellison, Amy Finkelstein, Drew Fudenberg, Paul Joskow, Whitney Newey, Pietro Ortoleva, Paul Oyer, Ariel Pakes, Alex Rees-Jones, Frank Schilbach, Chad Syverson, and Alex Wolitzky for helpful conversations, comments and suggestions. Oishi Banerjee, Andrew Komo, Manasi Kutwal, Angelo Marino and Jett Pettus provided invaluable research assistance. The authors acknowledge support from the Alfred P. Sloan Foundation (2022-17182), JPAL Healthcare Delivery Initiative, and MIT SHASS. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology
Nikhil Agarwal, Alex Moehring, Pranav Rajpurkar, and Tobias Salz
NBER Working Paper No. 31422
July 2023
JEL No. C50,C90,D47,D83

ABSTRACT

While Artificial Intelligence (AI) algorithms have achieved performance levels comparable to human experts on various predictive tasks, human experts can still access valuable contextual information not yet incorporated into AI predictions. Humans assisted by AI predictions could outperform both human-alone or AI-alone. We conduct an experiment with professional radiologists that varies the availability of AI assistance and contextual information to study the effectiveness of human-AI collaboration and to investigate how to optimize it. Our findings reveal that (i) providing AI predictions does not uniformly increase diagnostic quality, and (ii) providing contextual information does increase quality. Radiologists do not fully capitalize on the potential gains from AI assistance because of large deviations from the benchmark Bayesian model with correct belief updating. The observed errors in belief updating can be explained by radiologists' partially underweighting the AI's information relative to their own and not accounting for the correlation between their own information and AI predictions. In light of these biases, we design a collaborative system between radiologists and AI. Our results demonstrate that, unless the documented mistakes can be corrected, the optimal solution involves assigning cases either to humans or to AI, but rarely to a human assisted by AI.

Nikhil Agarwal
Department of Economics, E52-440
MIT
77 Massachusetts Avenue
Cambridge, MA 02139
and NBER
agarwaln@mit.edu

Alex Moehring
MIT
100 Main St
E62-390
Cambridge, MA 02472
moehring@mit.edu

Pranav Rajpurkar
Harvard Medical School
10 Shattuck St
Boston, MA 02115
United States
pranav_rajpurkar@hms.harvard.edu

Tobias Salz
Department of Economics, E52-460
MIT
Cambridge, MA 02139
and NBER
tsalz@mit.edu

Randomized controlled trials registry entry are available at
https://www.socialscienceregistry.org/trials/9620 and
https://www.socialscienceregistry.org/trials/8799

Appendix is available at http://www.nber.org/data-appendix/w31422

"We should stop training radiologists now. Its just completely obvious that within five years, deep learning is going to do better than radiologists."

– Geoffrey Hinton (in 2016)

1 Introduction

Artificial intelligence (AI) is a general-purpose technology with transformative potential similar to that of the steam engine and electricity (Brynjolfsson and Mitchell, 2017; Brynjolfsson et al., 2017; Agrawal et al., 2018; Acemoglu and Johnson, 2023; Goldfarb et al., 2023; Frank et al., 2019). But, in contrast to the transformation during the industrial revolutions, AI can potentially displace humans from tasks that require complex reasoning (Webb, 2019; Felten et al., 2019; Brynjolfsson and Mitchell, 2017). Advances in AI have therefore caused concern about the role of human work, even in highly skilled occupations (Ford, 2015). Indeed, a growing literature shows that AI tools based on machine learning can outperform humans on a host of predictive tasks, including those typically performed by experts (Liu et al. 2019; Lai et al. 2021; Mullainathan and Obermeyer 2019; Kleinberg et al. 2017).

However, Hinton's prediction about radiology — an iconic example in machine learning — may have been premature.[1] Many argue that instead of being replaced by AI, it is optimal to have human radiologists use AI assistance, at least for the foreseeable future (Langlotz, 2019; Agrawal et al., 2019). In addition to considerable legal and regulatory challenges that stand in the way of full automation, there may be potential gains from combining human expertise with AI input that cannot be realized by exclusively relying on one or the other. For example, radiologists may correct mistaken AI predictions or may have access to information about the clinical context that has not yet been systematized and built into AI tools. Current regulatory practice is consistent with these arguments – approved AI tools for clinical decision-making typically play a supporting role rather than operating autonomously (see Norden and Shah, 2022; Harvey and Gowda, 2020, for example). Similar reasons are likely to result in partial task automation or collaboration between human experts and AI being optimal in other contexts.

We run an experiment to examine how radiologists use AI predictions, whether AI assistance improves their performance, estimate a model of (potentially biased) belief updating, and analyze what this model implies about the optimal form of collaboration between AI and humans.[2] If humans correctly combine AI predictions with their own information, then AI

[1] A more nuanced but qualitatively similar prediction that machine learning tools will displace radiologists is conveyed in Obermeyer and Emanuel (2016).

[2] We will use the terms 'humans,' 'radiologists,' and 'participants' interchangeably.

assistance unambiguously improves prediction and decision quality. Unless the costs in terms of human effort outweigh these benefits, it is optimal to provide AI assistance to humans. However, substantial literature in economics suggests that humans may err when making probabilistic judgments by deviating from the benchmark model of Bayesian updating with correct beliefs (see Benjamin et al., 2019, for a review). The optimal approach for combining human and AI information in the presence of these mistakes is non-trivial.

The experiment employs professional radiologists whom we recruit through teleradiology companies to diagnose retrospective patient cases. We experimentally manipulate the information set radiologists have access to when making decisions in a two-by-two factorial design. In the minimal information environment, we provide only the chest X-ray image to which we add either AI predictions or contextual information, or both. The AI information treatment provides probabilities that a patient case is positive for a potential chest pathology generated using an algorithm trained on over 250,000 X-rays with corresponding disease labels (Irvin et al., 2019). This algorithm was shown to perform comparably to board-certified radiologists. The contextual information treatment provides clinical history information that radiologists typically have available but, for privacy reasons, was not available to train the AI. This information includes the treating doctors' indications, the patient's vitals, and the patient's labs.

We first estimate the treatment effects of those informational interventions on radiologists' prediction accuracy and the probability of making a correct decision. We then test alternative models of biased belief updating to investigate whether radiologists exhibit systematic deviations from a Bayesian benchmark when incorporating AI predictions. For example, humans may suffer from automation bias, a tendency to place more weight on machine-provided predictions than on one's own information.[3] Additionally, humans may treat AI predictions as independent of their own information (Enke and Zimmermann, 2019). Finally, we evaluate various forms of human-AI collaboration – in terms of diagnostic performance and costs of expert time – that decide, as a function of the AI prediction, to use only the AI input, only the human input, or the human input with access to AI.

Diagnostic radiology is an ideal laboratory for investigating human-AI interaction for various reasons. Deep learning has made significant advances in radiology, surpassing human decision-making in many cases, and such algorithms are already clinically deployed Rajpurkar and Lungren (2023). It, therefore, serves as a useful leading indicator for other professions

[3]This terminology is borrowed from the literature that dates to the proliferation of computerized automated support systems in aviation, which raised concerns about human complacency or automation bias (see Alberdi et al., 2009, for an overview).

where similar developments will follow. Moreover, radiology is a highly-paid medical specialty, which means that the potential benefits from productivity enhancements through AI are large. Radiology is also ideal from a research perspective. Unlike other physicians, radiologists do not have direct contact with patients, allowing us to run a decision experiment that resembles their normal workflow through a remote interface that we developed. To aid quantitative analysis, instead of obtaining a free-text report, we collect radiologist's assessed probability that a given chest pathology is present and a binary clinical recommendation of whether to treat or follow up for that pathology. For our experiment, we hired 180 radiologists through teleradiology companies that serve US hospitals and offer the services of both US-based and non-US-based radiologists.

Analyzing the quality of our participants' assessments requires establishing, for each patient case, a standard or ground truth. An important challenge in our setting is that definitive diagnostic tests do not exist for most thoracic pathologies and, even when they do exist, are selectively performed depending on a radiologist's recommendation. We therefore follow the machine learning literature (Sheng et al., 2008) and construct a diagnostic standard by aggregating the assessments of five board-certified radiologists at Mt. Sinai Hospital with at least ten years of experience and chest radiology as a sub-specialty. We assess the robustness of all our results by constructing a (leave-one-out) ground truth using the assessments of our experimental participants and by varying the aggregation method.

There are two important challenges in designing the experiment, which we address using a combination of experimental designs. First, while teleradiology firms allow us to recruit radiologists for our experiment we need to compensate them at the market rate, making data collection expensive. An across-participant design would therefore be cost-prohibitive except for extremely large effect sizes. In the first design, we, therefore, randomize each participant into an order in which they are exposed to the four informational environments but do not encounter cases repeatedly. Thus, this design allows for within-participant comparisons in addition to an initial pure across-participant comparison. The second challenge is that, to estimate a model of belief updating, we need to observe radiologists' assessments of a given case both with and without AI assistance. Moreover, the assessment without AI assistance is required to estimate the Bayesian benchmark, which we can compare to the assessment that incorporates AI assistance. Our second design addresses this issue by asking each participant to diagnose each patient case once in each of the four information environments in random order while ensuring at least a two-week wash-out period between two encounters of a given case. This wash-out period is intended to eliminate the memory of prior information and

anchoring effects.[4] Most of our treatment effect analysis pools data from the different designs whereas our model estimates of belief updating only uses data where a radiologist reads the same case both with and without AI assistance.

We find that AI assistance does not improve humans' diagnostic quality on average even though the AI predictions are more accurate than almost two-thirds of the participants in our experiment. Moreover, the zero average effect cannot be explained by the participants ignoring these predictions – we observe that radiologists' reported probabilities move significantly towards the AI's predictions with AI assistance. Instead, the zero effect of AI assistance is driven by heterogeneous treatment effects – diagnostic quality increases when the AI is confident (e.g. the predicted probability is close to zero or one) but decreases when the AI is uncertain. In parallel, AI assistance improves diagnostic quality for patient cases in which our participants are uncertain, but decreases quality for patient cases in which our participants are certain. In contrast, providing clinical history does improve diagnostic quality suggesting that humans have additional valuable information that has not yet been incorporated into AI predictions.

An upshot of the results is that information available only to radiologists is useful, but human experts do not correctly combine their information with AI predictions. Specifically, the result that AI assistance can reduce predictive performance cannot be rationalized if our participants are Bayesians with correct beliefs because the AI assistance provides weakly more information to the decision-maker.

Motivated by these results, we analyze two types of deviations from the benchmark model with correct updating to link errors in probabilistic judgement and optimal deployment of AI assistance.[5] The first type of deviation occurs when agents do not put correct weights on their own information and AI information. We describe this deviation using the approach introduced in Grether (1980; 1992) (see Benjamin, 2019, for a review) to define biases in belief updating. We say that an agent exhibits own-information bias if they over-weights their baseline information when provided with AI assistance and own-information neglect if they under-weights it. Analogously, an agent exhibits automation bias if they over-weights the AI information relative to their own and automation neglect if they under-weights it. The second type of deviation occurs if agents utilize an incorrect joint distribution of their

[4] To ensure that our results do not rely on the wash-out being successful, a third design obtains an assessment with AI assistance only after assessments without AI assistance have been obtained. However, this third treatment is subject to order effects. We find no evidence of order effects on diagnostic quality although there is evidence that familiarity with the interface increases the speed with which participants go through patient cases.

[5] We will remain agnostic about whether the deviations we consider are due to non-Bayesian updating or can be explained by Bayesian updating with incorrect prior beliefs.

own information and AI information – an example of such a deviation is correlation neglect (Enke and Zimmermann, 2019). Our theoretical analysis shows that if agents exhibit only automation neglect, then AI assistance unambiguously increases diagnostic quality. All other forms of biases we consider result in AI assistance reducing diagnostic quality for certain realizations of AI and own information.

We then use the data from our experiment to estimate empirical analogs of the deviations described above and select the model that best describes the treatment effects we document. This exercise requires us to solve several challenges unique to a naturalistic setting where the experimenter does not control the information structure. We find that in the model that best describes our data, agents exhibit automation neglect and act as if their own information and AI predictions are independent (conditional on the truth), even though this is not the case. Interestingly, we do not find evidence of substantial own-information bias or neglect.

Finally, we take our data to estimate the trade-off between diagnostic quality and radiologist time when AI predictions can be selectively utilized, and design an optimal human-AI collaborative system. For each pathology and conditional on the AI prediction for that pathology, our experiment allows us to compute both the diagnostic quality and time taken if a patient case is diagnosed by the AI, a randomly chosen radiologist without AI assistance, and a radiologist assisted with the AI prediction. We use these data to train a classification algorithm that yields the frontier of diagnostic quality and total radiologist time. The results from this exercise mirror our treatment effect analysis – because radiologists take more time with AI and do not correctly incorporate the AI's information, the majority of cases are optimally decided either by the radiologist or the AI alone but not by the radiologist with access to AI.

Related Literature

A growing body of literature in computer science has explored the predictive performance of humans versus machine learning algorithms, with radiology often serving as a key area of application (Rajpurkar et al., 2018, 2017). Additionally, the study of human-AI collaboration has become an increasingly important facet of medical AI research (Tschandl et al., 2020; Reverberi et al., 2022). For comprehensive overviews of these areas, see Rajpurkar et al. (2022); Hosny et al. (2018); Zhou et al. (2021); Lai et al. (2021). Research on the effectiveness of human-AI collaboration is evolving, with notable studies in radiology including Rajpurkar et al. (2020); Kim et al. (2020); Park et al. (2019); Seah et al. (2021); Fogliato et al. (2022). An active literature studies whether AI assistance benefits radiologists, and which radiologists benefit the most (Rajpurkar et al., 2020; Seah et al., 2021; Ahn et al., 2022; Sim et al., 2020; Gaube et al., 2023). In contrast to prior studies, we recruit a large group of high-

skilled experts from teleradiology companies under contracts that allow us to incentivize our participants. A key differentiating factor of our research is that, unlike previous studies which mainly concentrated on performance, our work emphasizes behavioral biases and their impact on human-AI interaction.

An emerging literature in economics also compares human and AI performance. Within economics, these studies tend to rely on observational approaches, with examples addressing issues in medicine (Ribers and Ullrich, 2022; Mullainathan and Obermeyer, 2019) and bail decisions (Kleinberg et al., 2015; Angelova et al., 2022). However, analyses based on observational data face critical identification challenges, such as the selective labels problem (see Kleinberg et al., 2017; Mullainathan and Obermeyer, 2019; Rambachan, 2021)). A limited set of studies use quasi-experimental approaches (e.g., Stevenson and Doleac, 2019; Angelova et al., 2022) or randomized controlled trials (e.g., Imai et al. (2020); Noy and Zhang (2023); Grimon et al. (2022)) to investigate human use of AI tools, typically focusing on overall performance or variability in participant response. We add to this literature by developing an experimental approach that manipulates the information environment that calculates and compares behavior with a Bayesian benchmark to document systematic biases and demonstrate that these biases lead to a non-trivial delegation problem.[6]

While several studies in behavioral economics have documented errors in probabilistic judgment and belief formation, they do not focus on the issues surrounding human-AI interaction (c.f. Tversky and Kahneman, 1974; Benjamin et al., 2019; Enke and Zimmermann, 2019; Conlon et al., 2022, for example). Our definitions of own-information and automation bias build on the framework in Grether (1980). We contribute to this literature in two ways. First, we develop an approach to estimate the parameters of the model in Grether (1992) in an environment where the joint distribution of the signals cannot be controlled (or partialled out) by the researcher.[7] This approach is necessary because we cannot modify the signal within medical images. Second, we link the design of AI information provision to the (biased) updating rule that humans use. This link shows that the use of AI information by humans is an important and practical application of the ideas in this literature.

[6]Our finding that radiologists exhibit automation neglect is related to those in Dietvorst et al. (2015), which shows that humans are averse to following algorithmic recommendations as compared to human recommendations. This aversion can be reduced if humans are allowed to modify the algorithm's recommendation (Dietvorst et al., 2018).

[7]Most applications that we are aware of rely on one of two experimental approaches. In the first approach, the researcher can partial out either the prior information or the likelihood ratio of the signal provided, for example in the classic bookbag-and-poker-chip experiments (see Benjamin et al. 2019; Benjamin 2019, for reviews). In the second approach, the researcher directly provides signals from a known joint distribution (e.g. Conlon et al., 2022).

Our analysis of optimal human-AI collaboration is related to papers that build delegation algorithms to predict the types of cases for which human performance exceeds machine performance (e.g. Mozannar and Sontag, 2020; Raghu et al., 2019; Bansal et al., 2021). Relative to this work, our analysis uses a decision-theoretic model and specific human biases to trace their consequences for optimal AI deployment.

Finally, our work also adds to the literature on decision-making in the health care context (e.g. Abaluck et al., 2016; Currie and MacLeod, 2017; Gruber et al., 2021; Chan et al., 2022; Chandra and Staiger, 2020). This work aims to understand predictions and payoffs from observational data on medical decisions, objectives that are achievable under less stringent functional form restrictions in our experimental approach. An important distinguishing feature is that none of these papers consider the effects of AI predictions on medical decision-making.

Overview

The rest of the paper is organized in the following way. Section 2 introduces our model of a decision-maker in a diagnostic setting. Section 3 describes the necessary details of the setting and our experimental design. Section 4 discusses the treatment effects. Section 5 estimates a descriptive model of deviations from Bayesian updating. Section 6 shows the gains achievable under the optimal collaboration between radiologists and AI.

2 Conceptual Model

Our study focuses on classification problems and, specifically machine learning algorithms within the domain of AI tools. These algorithms are designed to predict the appropriate classification for a given case and may assist a human decision-maker. This decision-maker, indexed by r, must take a binary action $a_{ir} \in \{0,1\}$ on case i based on a prediction of a binary state $\omega_i \in \{0,1\}$. The realized payoff $u_r(a_{ir}, \omega_i)$ from an action depends both on the state and the action. The expert does not know ω_i but observes, depending on the information environment, a subset of two signals that are potentially informative about the state. The first signal is generated by a prediction algorithm (AI), with realizations $s_i^A \in S^A$. The second signal is directly obtained by the expert, with a realization $s_{ir}^E \in S^E$. These signals are of arbitrary dimension. The joint distributions of the signals conditional on the state is given by $\pi_r(\cdot|\omega) \in \Delta\left(S^A, S^E\right)$, with prior probabilities over the state $\pi(\omega)$. We do not place any restrictions on $\pi_r(\cdot|\omega)$. The signals need not be independent conditional on the state of the world, and the signal distribution may depend on the expert, capturing skill (Chan et al., 2022). We also allow for cases when one of the signals is more informative than

the other (Blackwell, 1953).

Assume that the human's objective is to match the action to the state. Thus, the human faces a classification problem. It is without loss of generality to normalize the payoff from taking the action that matches the state to zero. Let $c_{FP,r}$ be the disutility of human r if she takes the action $a = 1$ when the state is $\omega = 0$ (false positive) and $c_{FN,r}$ be the disutility if she takes the action $a = 0$ when the state is $\omega = 1$ (false negative). The payoff of the human is therefore

$$u_r(a, \omega) = -1 \cdot \{a = 1, \omega = 0\} \cdot c_{FP,r} - 1 \cdot \{a = 0, \omega = 1\} \cdot c_{FN,r}. \tag{1}$$

We allow the human's posterior belief given the observed signals to deviate from those implied by the true probability law $\pi_r(\cdot|\omega)$. Specifically, let $s_{ir} \subset \left\{s_r^A, s_{ir}^E\right\}$ be the subset of signal realizations observed by the human r and $p_r(\omega \,|s_{ir}) \in [0, 1]$ be the human's posterior belief that the state is $\omega \in \{0, 1\}$ when she observes s_{ir}. Suppressing the dependence of signals on the pair (i, r), the human's action given the signal s is

$$a_r^*(s; p_r) = 1 \cdot \left\{ \frac{p_r(\omega = 1 \,|s)}{p_r(\omega = 0 \,|s)} > c_{rel,r} \equiv \frac{c_{FP,r}}{c_{FN,r}} \right\}. \tag{2}$$

The expected payoff from following $a^*(s)$ is

$$V_r(s; p_r) = E\left[u_r\left(a_r^*(s; p_r), \omega\right) |s\right] = \sum_{\omega} u\left(a_r^*(s; p_r), \omega\right) \pi_r(\omega \,|s),$$

where decisions are based on the human's belief p_r, but are evaluated according to the true law π_r. Because we allow for p_r to differ from π_r, the odds ratio $\frac{p_r(\omega=1|s)}{p_r(\omega=0|s)}$ may differ from the analogous quantity constructed using π_r. Thus, the action $a_r^*(s; p_r)$ can deviate from the optimal action $a_r^*(s; \pi_r)$ given the signal $s = \left(s^A, s^E\right)$. Except in knife-edge cases, $V_r(s; p_r)$ is lower than $V(s; \pi_r)$ whenever $a^*(s; p_r) \neq a^*(s; \pi_r)$.

A key objective is to analyze deviations in humans' use of AI signals from the benchmark model of Bayesian updating with correct beliefs (about the joint distribution of the signals and the state). Bayes' rule implies that, given the signals $\left(s_i^A, s_{ir}^E\right)$, the log-odds of $\omega = 1$ to $\omega = 0$–the key decision-relevant quantity–is given by

$$\log \frac{\pi_r\left(\omega_i = 1 \left| s_i^A, s_{ir}^E \right.\right)}{\pi_r\left(\omega_i = 0 \,|s_i^A, s_{ir}^E\right)} = \log \frac{\pi_r\left(s_i^A \left| \omega_i = 1, s_{ir}^E \right.\right)}{\pi_r\left(s_i^A \,|\omega_i = 0, s_{ir}^E\right)} + \log \frac{\pi_r\left(\omega_i = 1 \left| s_{ir}^E \right.\right)}{\pi_r\left(\omega_i = 0 \,|s_{ir}^E\right)}. \tag{3}$$

The second term on the right-hand side is the posterior log-odds ratio for the two states

$\omega_i = 1$ to $\omega_i = 0$ given that the human's signal is s_{ir}^E. The first term is positive if, given the realization s_{ir}^E, the signal s_i^A is more likely if $\omega_i = 1$ as compared to $\omega_i = 0$. In this case, the posterior odds shift in favor of the state $\omega_i = 1$. Analogously, the odds shift away from $\omega_i = 1$ if, given s_{ir}^E, the signal s_i^A is more likely if the state were $\omega_i = 0$.

The task of empirically analyzing deviations in participants' beliefs from the benchmark in equation (3) is challenging because the signals differ across cases i and humans may, due to differences in skill, have heterogeneous signal distributions $\pi_r(\cdot)$. The ideal dataset would elicit $p_r\left(\omega_i = 1 \left| s_i^A, s_{ir}^E \right.\right)$ and $\pi_r\left(\omega_i = 1 \left| s_{ir}^E \right.\right)$ for the same case and the same human to keep the signals $\left(s_i^A, s_{ir}^E\right)$, and human-specific parameters fixed, and use the latter to estimate $\pi_r\left(\omega_i = 1 \left| s_i^A, s_{ir}^E \right.\right)$ using equation (3).

However, empirically implementing this strategy requires us to confront several experimental and methodological challenges. First, the experiment needs to be sensitive to anchoring and order effects when eliciting $p_r\left(\omega_i = 1 \left| s_i^A, s_{ir}^E \right.\right)$ and $\pi_r\left(\omega_i = 1 \left| s_{ir}^E \right.\right)$ from participants. Second, the potential for measurement error and the dependence of the first term in equation (3) on a latent signal s_{ir}^E complicates the exercise. We defer our approach to this second challenge to section 5, and turn our attention to the experimental design.

3 Setting and Experiment

Our experiment elicits the probability of a pathology's presence $p_r(\omega = 1 \,|s)$ and a recommended treatment/follow-up decision a under varying information treatments. There are four information treatments in the experiment, with only the chest X-ray in the minimal information case, to which we add AI assistance, contextual information, or both. Next, we describe the experimental context and interface before presenting the design of our experiments.

3.1 Experimental Context

3.1.1 Radiology

Radiologists diagnose the presence of a given pathology at the request of a treating physician. The information available to a radiologist consists of diagnostic images (e.g. chest X-rays), any relevant medical history (e.g. laboratory results), and clinical indication notes of the treating physician.[8] The treating physician's notes are of varying detail levels – they may provide no clinical information or guidance, request the analysis of a specific pathology, or

[8]Radiologists are rarely in direct contact with the patient or the treating physician, except via the formal information exchange outlined here.

only list the patient's primary symptom (see appendix B.2 for examples). Irrespective of the pathology suspected by the treating physicians, radiologists are expected to report all pathological findings.

Because image-based classification is a core task performed by radiologists, a high-paying profession, it is not surprising that AI tools have made significant inroads in the field in the last decade. Recent advances in deep learning methods for image recognition have yielded algorithms that can match or surpass the performance of human radiologists (Obermeyer and Emanuel, 2016; Langlotz, 2019). As of 2020, 55 companies offered a total of 119 algorithmic products of which 46 have FDA approval (Tadavarthi et al., 2020). Most products related to clinical decision-making are marketed as support tools as opposed to autonomous tools, partly due to regulatory and liability issues (Harvey and Gowda, 2020).

3.1.2 CheXpert

We provide AI assistance using the CheXpert model, which is a deep learning prediction algorithm for chest X-ray interpretation (Irvin et al., 2019). This model is trained on a dataset of 224,316 chest radiographs of 65,240 patients labeled for the presence of fourteen common chest radiographic pathologies.[9] The algorithm does not use any other patient information, such as the clinical history or vitals.[10] Nonetheless, a prior version of this algorithm was shown to match or surpass the performance of board-certified radiologists from Stanford Hospital on five pathologies (Patel et al., 2019). These study results are also presented to our participants when introducing the AI tool. Section 4 confirms that the algorithm outperforms a majority of radiologists in our experiment. We relegate additional details about the algorithm to appendix B.3. The algorithm assistance to our participants will be in the form of a vector of probabilities for the presence of every pathology.[11]

[9]The term artificial intelligence is typically reserved for a system of different prediction tasks to mimic a more complex set of behaviors, whereas machine learning is concerned with one specific prediction task. For a detailed discussion of this distinction see, for instance, (Taddy, 2018).

[10]While large datasets of images are increasingly available (e.g. Kramer et al. 2011, Johnson et al. 2016, Irvin et al. 2019) it is significantly more difficult to construct such datasets for other patient information due to the compulsory manual review of textual data for HIPPA compliance.

[11]Some algorithms attempt to make their predictions explainable to a human by highlighting the parts of the image that drive a specific prediction. However, prior studies show that providing such localization in addition to the numeric output does not improve the accuracy of radiologists (Gaube et al., 2022). A quantitative output allows us to compute a Bayesian benchmark to the radiologist's prediction, which is otherwise difficult.

3.2 Experimental Designs

Our experiment varies the information available to diagnose patient cases—participants may or may not receive AI assistance and access to the clinical history. The X-ray is shown under all information conditions. We expose our participants to all four possible information conditions: X-ray only, henceforth *XO*; clinical history without AI, henceforth *CH*; AI without clinical history, henceforth *AI*; and both clinical history and AI, henceforth *AI+CH*.

There are two objectives of our experiment. The first is to compute the treatment effects of AI and CH on diagnostic quality and radiologist time. The second is to analyze systematic deviations from the Bayesian benchmark.

Both these objectives are complicated by the likely heterogeneity in radiologist skills. For estimating treatment effects, radiologist heterogeneity implies that a design that randomizes treatments only across radiologists will require a large participant pool except for extremely large effect sizes. Our participants are highly paid experts, making this approach expensive. And, as explained in section 2, across-radiologist variation in information treatments is not tailored for the second objective. We would ideally know how a given radiologist changes her assessment for the same case under a different information condition.

Our approach to address these challenges is to use a combination of three different experimental designs, each with certain advantages and disadvantages. Appendix B.1 illustrates the three design variations.

3.2.1 Design 1 (Figure B.1)

In the first design, participants are assigned to a random sequence of the four information treatments. Each information condition is assigned fifteen cases at random without repetition. Participants read all 15 cases in one information environment before moving to the next one.

This design builds in both across- and within-participant variation in information treatments. The within-participant variation has greater power because it controls for participant heterogeneity at the potential cost of order effects. The concern of order effects is both testable and mitigated by the randomization of treatment sequence across subjects.

This first design is well-suited to estimate treatment effects of our information environments. However, as mentioned earlier, it is not ideal for estimating an empirical analog to equation (5) because no case is encountered twice.

3.2.2 Design 2 (Figure B.2)

Radiologists diagnose each patient case in each of the four information environments in the second design. For the moment, set aside concerns arising from the feature that the same radiologist encounters the same case multiple times. This design will allow us to estimate an empirical analog to equation (5). It also has the added benefit of controlling for both case-radiologist heterogeneity because, unlike in the previous design, we can conduct within-case-radiologist comparisons across treatments.

Because radiologists repeatedly encounter cases, we need to address the potential for order effects due to memory. For example, radiologists might anchor on their previous assessment using AI predictions or contextual information and might remember this information the next time the same case is encountered. We, therefore, limit radiologists' ability to remember either their diagnosis or previously provided information by using a "washout" interval between two encounters of the same case.[12] Specifically, radiologists complete the experiment in four sessions that are separated by at least two weeks. Each session is similar to the first design: radiologists diagnose fifteen cases in each of the four information environments with no case repeated within a session. Across sessions, the information environment under which a given case is diagnosed is permuted. Thus, by the end of the fourth session, each of the sixty cases is diagnosed exactly once in each information environment. Our results are consistent with the washout being effective – radiologists' predictions do not move towards the AI prediction if it was provided in a prior session but do if it is provided in the current session (see figure C.28).

3.2.3 Design 3 (Figure B.3)

In the third design, we address residual concerns about the order effects of radiologists diagnosing cases with AI before those without AI–whether due to anchoring, memory, or experimenter demand–by having participants diagnose fifty cases, first without and then with AI assistance. Within each block, clinical history is randomly provided in either the first or second half of images.

This design also allows us to conduct within case-radiologist comparisons. The potential disadvantage of this design is that we cannot distinguish order effects from the effect of providing AI. This issue is unavoidable given the guiding principle that participants receive weakly more information about a case during a repeat encounter. However, we can test for

[12] This principle has been used in computer science (Seah et al., 2021; Conant et al., 2019; Pacilè et al., 2020).

and do rule out order effects on accuracy based on the first two designs.

3.2.4 Participant Recruitment

Participants for the first and third designs, which constitute the majority, were recruited through teleradiology companies. The teleradiology companies allow us to recruit several experts in a relatively liquid spot market, a practice that is now common for decision experiments with non-expert subjects (Hunt and Scheetz, 2019). Most healthcare providers in the US rely on these companies' services, although many large hospitals have on-call radiologists (Rosenkrantz et al., 2019). We work with teleradiology companies that serve US hospitals and offer the services of both US-based and non-US-based radiologists. Our contracts with teleradiology companies specify a piece-rate, and the companies, in turn, compensate the participants with a piece-rate.[13] In addition, we provided monetary incentives for accuracy to a subset of radiologists, as described in the next section.

The second design required us to work with a partner who could guarantee subjects' participation over several months. We collaborated with VinMac healthcare system in Vietnam to recruit their staff radiologists to ensure continued participation. VinMac is in the process of developing its own in-house AI capabilities and was willing to assist with our experiment in exchange for recognition in a publication of the resulting dataset. The VinMac radiologists did not receive receive any payments to participate in the experiment but we find that their perfomance is very close to the performance of the tele-radiologists.

In total, 180 radiologists participated in our experiment. Close to 25% of our participants are US-based, 20% have a degree from a US institution, 80% are affiliated with a large clinic, and 61% with an academic institution. As demonstrated in appendix C.3, the quality of the assessments made by the radiologists in our study is comparable to that of the staff radiologists from Stanford University Hospital, who originally diagnosed the patient case.

3.2.5 Incentives

We cross-randomize incentives for accuracy in the first and third designs but not the second because of the specific ways in which our partner's radiologists are employed.[14] Payments were determined following the binarized scoring rule in Hossain and Okui (2013), where truth

[13] We worked with three companies with piece-rates ranging from $7.50 to $13.00.

[14] There does not appear to be a consensus in the experimental literature on whether incentives are superior to non-incentivized responses when eliciting beliefs (see Danz, Vesterlund, and Wilson, 2020; Charness, Gneezy, and Rasocha, 2021). A comparison between the incentivized and non-incentivized participants allows us to answer this question in the context of this experiment.

is determined as described in section 3.3.1 below. This incentive scheme uses a loss function of the mean squared prediction error, averaging over patient cases and pathologies, and the respondents earn a fixed bonus of $120 if a random draw is less than the loss function. This bonus is more than 20% of the base payment to teleradiology firms. We explain to the participants that expected payments are maximized if they provide their best estimates using a non-mathematical description of the payment rule. We specify the distribution so that 30% of pilot participants would earn the bonus, cross-randomized with the other two treatment arms.

3.3 Implementation and Data Collection

3.3.1 Patient Cases and Ground Truth

The experiment uses 324 historical patient cases with potential thoracic pathologies from Stanford University's healthcare system. For each case, we have access to the chest X-ray and the clinical history in the form of the primary provider's written notes, the patient vitals, and demographics.[15] The use of retrospective cases allows us to avoid ethical and other issues that would arise when experimenting in high-stakes settings.

Our analysis requires constructing ω_i for each patient case. There are important challenges in using an observational dataset of patient health records to construct this field. One approach would be to use the results from further medical tests. Unfortunately, definitive gold-standard tests do not exist for most thoracic pathologies.[16] Even when follow-up tests are conducted, they are selected on the likely presence of a pathology, an issue referred to as the selective labels problem (e.g. Mullainathan and Obermeyer, 2019). Medical outcomes from patient health records also do not suffice because actions taken by the treating physician in response to the radiology report contaminate these measures. Recent literature has suggested instrumental variables approaches for solving this selective labels problem, but this work targets population quantities and not a "ground truth" on each case (e.g. Chan et al., 2022; Mullainathan and Obermeyer, 2019).

We construct ω_i by aggregating the assessment of a group of expert radiologists, an approach common in computer science (Sheng et al., 2008; Mccluskey et al., 2021). Specifically, we

[15] All cases are first encounters with no prior X-ray as a comparison. We started with 500 cases that fit these primary criteria. We omitted pediatric cases from this set. Finally, a radiologist reviewed the cases to remove instances with poor image quality. The clinical history was manually reviewed to remove patient-identifiable information and cleared for public release.

[16] Many pathologies do not have commonly used non-imaging-based diagnostic tools. For instance, the presence of cardiomegaly – an enlarged heart – can only be determined using imaging tools, thoracic surgery, or an autopsy.

ask five board certified radiologists from Mount Sinai to read each of the 324 cases using the interface described above with the available X-ray and clinical history. For each case i and expert radiologist r, we, therefore, obtain $\pi_r\left(\omega_i = 1 \left| s^E_{i,r}\right.\right)$ for each pathology, which we aggregate to generate ω_i. Specifically, we classify $\omega_i = 1$ if $\sum_r \pi_r\left(\omega_i = 1 \left| s^E_{i,r}\right.\right)/5 > 0.5$. This approach immediately addresses the selective labels problem because the availability of assessments is not selected on the likelihood of a pathology being present. Results in Wallsten and Diederich (2001) suggest that, under weak conditions that allow for measurement error in the reports and correlations across reports, the aggregate opinion of several experts is highly diagnostic as long as the experts are median unbiased. To assess robustness, we aggregate $\pi_r\left(\omega_i = 1 \left| s^E_{i,r}\right.\right)$ using both a log-odds average and a leave-one-out mean of the assessments of all the radiologists in our experiment in the clinical history treatment condition.

3.3.2 Experimental Interface and Data Collected

We developed the interface to present the patient cases and to collect radiologists' predictions and decisions in collaboration with board certified radiologists at Stanford University Hospital and Mt. Sinai Hospital. In contrast to free-text reports, we designed it to generate structured and quantitative data on chest X-ray responses that resemble a typical radiological report. A guiding principle in the design was to mimic clinical practice and to present and obtain all clinically relevant information. We briefly describe this interface and provide images and further details in appendix B.

On the landing page of each case, a high-resolution image of a patient's X-ray is presented to the radiologist, with the functionality to zoom and adjust brightness and contrast. When the experiment calls to show the clinical history, the interface presents clinical notes, vitals, and laboratory results available at the time the X-ray was originally ordered. If the experiment provides AI assistance, participants are shown AI predictions for fourteen pathologies and a composite prediction for whether or not there are any relevant findings.

The data entry interface collects a radiologist's assessments of the probability that various thoracic pathologies are present for a given case. The probability that a pathology is present given the available information, i.e. $p_r\left(\omega = 1 \left| s\right.\right)$, is elicited using a continuous slider. We visually subdivide possible responses into five intervals with standard language labels used in written radiological reports to aid the participants.[17] Our radiology collaborators grouped pathologies into eight exclusive parent categories based on their type. Each group has children

[17] The specific labels are *"Not present","Very Likely", "Unlikely", "Possible", "Likely", and "Highly Likely".* Several radiological publications have suggested such standardized language for radiological reports. See for instance Panicek and Hricak (2016).

that are more specific, which may be further subdivided in some cases. The groups all correspond to a standard class of pathologies. For instance, "airspace opacity" is distinct from a "cardiomediastinal abnormality." In the main text we focus on the parent categories with AI predictions and drop further subdivisions from the analysis. We refer to those as top-level pathologies. Our results are robust to including the lower-level pathologies in the analysis as we show in appendix C.4. The interface categorizes thoracic pathologies into groups by type to ease data entry. For example, allowing the user to simultaneously set the assessed probability of each disease in a specific category to zero.[18] In addition, we elicited an overall bottom-line assessment of whether the radiologists considers the case normal or not.

We also ask for a binary "treatment/follow-up" recommendation for each pathology that is not definitively ruled out.[19] We will interpret this input as $a_r^*(s)$. In a real clinical setting, a recommendation to follow-up could trigger the treating physician to prescribe additional medical tests or interventions with potential costs and benefits. Thus, an optimal recommendation trades off the cost of false positives and false negatives when recommending an action as in section 2. The probabilistic assessments with the follow-up decision will allow us to estimate radiologists' relative cost of false positives and false negatives.

In addition to $p_r(\omega = 1 \,|\, s)$ and $a_r^*(s)$, we record active time, response times, and any click-stream data that results from the interaction with the interface. The participants are not explicitly informed about this monitoring, and there are no explicit time limits. Our experiment runs remotely, and participants connect to a server, which hosts the interface and records responses. The interface has been extensively tested beforehand on different browsers by conducting pilots with every company we recruited the participants from.[20]

[18] Our radiology collaborators grouped pathologies into eight exclusive parent categories based on their type. Each group has children that are more specific, which may be further subdivided in some cases. The groups all correspond to a standard class of pathologies. For instance, an "airspace opacity" is distinct from a "cardiomediastinal abnormality." The more specific disease of "bacterial pneumonia" manifests as an airspace opacity. This structure reduced the burden on our participants, and we piloted the interface with several radiologists specializing in the interpretation of chest X-rays. Prior clinical research on AI in chest X-Ray image classification has used similar hierarchies (see Seah et al., 2021, for example).

[19] The binary treatment/follow-up decision is only asked for pathologies where a follow-up is clinically relevant. This includes all pathologies with AI assistance.

[20] This interface is browser-based and built using the o-tree framework Chen et al. (2016). Since we are not directly communicating with our participants we also deploy a device fingerprinting service from fingerprint.com to ensure that there are no repeat participants.

3.3.3 Participant Training

We train the participants using a combination of written instructions and a video. The materials provide an overview of the experimental tasks, the interface, and information about the AI assistance tool. The firms and the participants know that the research study involves retrospective patient cases. To train participants on the AI tool, we provide them with materials that explain the development of the algorithm, present metrics of its performance on various diseases, and summarize the algorithm's performance relative to radiologists based on prior research. The participants are informed that the algorithm only uses the chest X-ray to form predictions, and this knowledge is later tested in a comprehension question. In addition, we show the participants fifty example cases that show the X-ray and clinical history next to the AI output. After the instructions, they answer eight comprehension questions, which the participants must answer correctly before proceeding to the experiment. We also include an endline survey. We do not directly interact with the subjects except to field questions about the experiment or provide tech support.[21] The complete set of instructions is provided in appendix B.2.

4 Estimated Treatment Effects

4.1 Overall Performance of AI and Radiologists

This section's analysis focuses on measures of performance (deviation from ground truth, incorrect decision), deviation from AI prediction, and measures of effort. Table 1 summarizes the data on these measures and sample sizes from our experiment. The main text focuses on top-level pathologies with AI with robustness to other pathology groups relegated to appendix C.4.

Radiologists make the correct follow-up/treatment recommendation on approximately 70% of case-pathologies on average and spend ~2.8 minutes per case with large variability across cases. All summary statistics are very similar across the three expertimental designs (tables C.7 and C.9). For instance, the average deviation from the ground truth for the three designs ranges from 0.191 to 0.232, and average active time ranges from 2.58 to 3.03 minutes.

[21] We are blinded to the participants' treatment status while the experiment is in progress.

Table 1: Summary Statistics

	All Designs		**Design 1**	
	Mean	SD	Mean	SD
	(1)	(2)	(3)	(4)
Reported Probability	0.233	0.290	0.214	0.287
Decision	0.322	0.467	0.277	0.448
Deviation from Ground Truth	0.223	0.281	0.218	0.290
Deviation from AI	0.191	0.169	0.201	0.171
Correct Decision	0.695	0.460	0.736	0.441
Active time	2.82	2.63	3.03	3.17
Observations	36,280		13,440	
Radiologists	180		112	

Note: This table presents summary statistics of the experimental data. Columns (1) and (2) present the mean and standard deviation for all designs while Columns (3) and (4) present the same for design 1 only. Decision is an indicator for whether treatment/follow-up is recommended, correct decision is an indicator for whether the decision matches the ground truth, deviation from ground truth is the absolute difference between the reported probability and the ground truth, deviation from AI is the absolute difference between the expert's reported probability and the AI's reported probability, active time is measured in minutes.

In a comparison study with three radiologists, Irvin et al. (2019) show that the CheXpert model yields a better classifier than two out of three radiologists on five pathologies and all three on a subset of three pathologies. Our results may differ from that because we use a different pool of radiologists, a different sample of cases, and reads with contextual information (clinical history) to construct the ground truth. The latter two differences raise the bar for the AI because they reflect differences in the data-generating process.

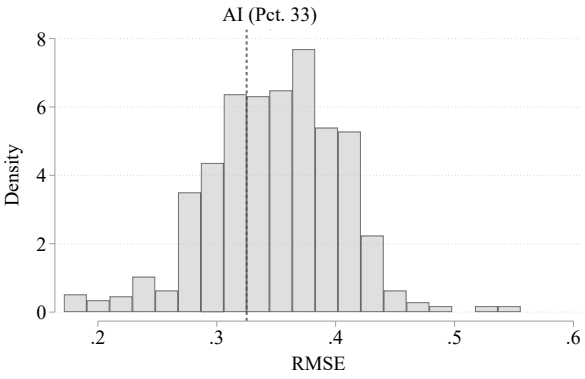
To benchmark the quality of AI predictions in our sample and the radiologists in our experiment, we compare our participant pool with the AI input using two performance measures. The first measure is derived from the receiver operating characteristic (ROC) curve, which measures the trade-off between the false positive and the true positive rate of a classifier as the cutoff for classifying a case as positive or negative is varied. It uses only ordinal information about the AI. A classifier that is not better than random has an AUROC value of 0.5 whereas a perfectly predictive classifier has a value of one. The second measure is the root mean squared error (RMSE), that utilizes cardinal information about the AI prediction. A

lower RMSE indicates higher performance.[22] We pool the data for top-level pathologies with AI for each radiologist's reports and for the AI's prediction.

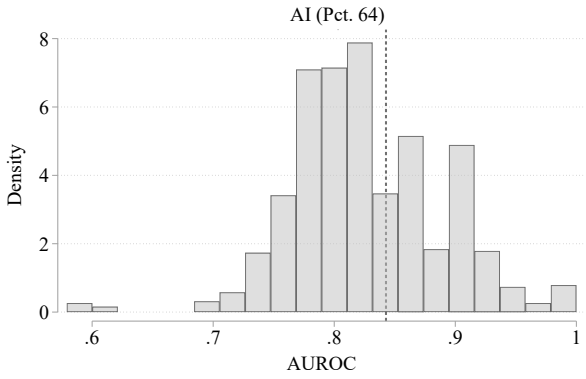
Figure 1 shows significant heterogeneity in performance across radiologists as well as the scope for AI assistance to improve radiologist performance. The AUROC shows that the bottom tail of the distribution of radiologists performs worse than the AI, whereas the upper tail predicts close to perfectly. This heterogeneity aligns with findings from observational data based on only treatment decisions (see Chan et al., 2022). The AI is more predictive than approximately two-thirds of radiologists, whether compared using AUROC or RMSE. Thus, there is ample room for AI assistance to improve the performance of radiologists. In fact, a majority of radiologists would do better on average by simply following the AI prediction. We present pathology-specific comparison results in appendix C.2.

Figure 1: Comparing AI performance to radiologists

(a) RMSE Radiologists and AI



(b) AUROC Radiologists and AI



Note: These histograms show distributions of two different accuracy measures of radiologist assessments alongside the AI's accuracy. The left graph shows the distribution of the RMSE while the right shows the distribution of the average AUROC. Both distributions are shrunk to the grand mean using empircal Bayes. These measures are for each radiologist and include the top-level pathologies. The dotted line is the average measure of the AI algorithm for the corresponding distribution. Only the assessments where contextual history information is available for the radiologists but not the AI prediction are considered. Robustness by design and ground truth definition can be found in sections C.4.1 & C.4.2.

We also compare the performances of our participants and the radiologist who originally diagnosed each patient case in appendix C.3.[23] There is no discernible difference between

[22] An important distinction between the two measures is that the AUROC is ordinal whereas RMSE is cardinal.

[23] We classified the original free text radiology reports associated with each case as positive, negative, or uncertain for each pathology using the CheXbert algorithm described in Smit et al. (2020). To facilitate comparisons, we also discretized the probability assessments elicited during the experiment into these three categorized. Then, we compared the accuracy of the original reads against the radiologists participating in the experiment.

the two groups, which is consistent with the hypothesis that radiologists participating in the study were of similar skill and exerted similar effort as the radiologists completing the original reads.

4.2 How do Radiologists Respond to AI and Contextual Information?

We now describe the effects of our information treatments estimated using the following specification:

$$Y_{irt} = \gamma_{h_i} + \gamma_{CH} \cdot d_{CH}(t) + \gamma_{AI} \cdot d_{AI}(t) + \gamma_{AI \times CH} \cdot d_{CH}(t) \cdot d_{AI}(t) + \varepsilon_{irt},$$

where Y_{rit} is an outcome variable of interest for radiologist r diagnosing patient case-pathology i and treatment t, and γ_{h_i} are pathology fixed effects since there are multiple pathologies h_i for each case in this pooled analysis. Treatments t vary by whether or not clinical history is provided $d_{CH}(t) \in \{0, 1\}$ and whether or not AI information is provided $d_{AI}(t) \in \{0, 1\}$. We report two-way clustered standard errors at the radiologist and patient-case level. The specification omits radiologist-specific fixed effects because the treatments are balanced within radiologists. Cases are also balanced across treatments (see appendix C.1), which suggests that case randomization was successful. We will also compute conditional treatment effects given ranges of the AI signal s_i^A that are grouped based on $\pi\left(\omega_i = 1 | \, s_i^A\right)$.

Table 2: Average treatment effects

	Deviation from AI		**Deviation from Ground Truth**		**Effort Measures**			
	All Designs	Design 1	All Designs	Design 1	All Designs		Design 1	
					Active Time	Clicks	Active Time	Clicks
Treatment	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
AI × CH	0.001	−0.001	0.004	0.003	−0.86	0.22	0.04	0.32
	(0.003)	(0.006)	(0.005)	(0.012)	(3.68)	(0.82)	(6.94)	(1.63)
AI	−0.038	−0.036	0.002	0.002	6.75	1.34	6.26	1.53
	(0.004)	(0.006)	(0.004)	(0.008)	(2.47)	(0.56)	(4.71)	(1.05)
CH	−0.001	−0.004	−0.011	−0.015	7.37	0.16	6.53	0.09
	(0.002)	(0.004)	(0.004)	(0.008)	(2.53)	(0.55)	(4.78)	(1.08)
Control Mean	0.211	0.221	0.227	0.224	157.57	44.25	161.56	40.71
	(0.007)	(0.008)	(0.010)	(0.011)	(4.46)	(1.21)	(7.05)	(1.53)
Pathology FE	Yes	Yes	Yes	Yes	-	-	-	-
Observations	36280	13440	36280	13440	14635	14635	6718	6718

Note: This table summarizes the average treatment effects (ATE) of different information environments on the absolute value of the difference between the radiologist probability and AI probability (column (1) and (2)); absolute value of the difference between the radiologist probability and the ground truth (columns (3) and (4)); and radiologists' effort measured in terms of active time and clicks (columns (5), (6), (7) and (8)). We either pool across all designs (All Designs) or condition on only design 1. Results on effort measure excludes five patient-cases with unaccounted time measure, and observations from design 3 because of learning effects in this set-up. The results are for the two top-level pathologies, airspace opacity and cardiomediastinal abnormality. Standard errors are two-way clustered at the radiologist and patient-case level in parenthesis. Robustness by design can be found in section C.4.1.

The above analysis focuses on the effects on the marginal distributions of the outcome variables Y_{irt} for each pathology. Thus, the specification abstracts away from interactions between pathologies in the effects on information provision, for example due to potential dependence between the predictions and decisions. In section 5, we will present evidence showing that the best fitting model has radiologists updating their beliefs as-if they do not account for dependence between pathologies.

The treatment effect analysis in the main text pools the three experimental designs and does not condition on the sequence in which subjects encounter information treatments. Appendix C.4.5 shows that our results are robust to including controls for order effects. The estimates from all three designs are similar to each other. They are also statistically indistinguishable from those that use only an across participant comparison from the first treatment encountered in design 1 (appendix C.4.3).

4.2.1 Do Radiologists Utilize AI Predictions?

We begin by testing whether radiologists respond to the information that the AI provides. The left panel in table 2 shows how the different information environments affect the disagreement of the radiologists' report with the AI's assessment, which is defined as $Y_{irt} = \left|p_r\left(\omega_i = 1 \middle| s_{irt}\right) - \pi\left(\omega_i = 1 \middle| s_i^A\right)\right|$. The term $p_r\left(\omega_i = 1 \middle| s_{irt}\right)$ is elicited whereas $\pi\left(\omega_i = 1 \middle| s_i^A\right)$ is the AI's predicted probability that $\omega_i = 1$. When t indicates that AI assistance is provided, then $s_{irt} = \left(s_{ir}^E, s_i^A\right)$, and when t indicates that AI assistance is not provided, then $s_{irt} = s_{ir}^E$. The signal s^E also depends on whether contextual information is provided.

The results show that radiologists utilize AI assistance: their predictions are closer to the AI predictions when provided with assistance. To see this, observe that the control means for the deviation from the AI are approximately 0.21 for both when we pool designs and for design 1 only. Treatments where AI is provided reduce this baseline average deviation by 18% (all designs) and 16% (design 1). ($p < 0.01$).

4.2.2 Treatment Effects on Diagnostic Performance

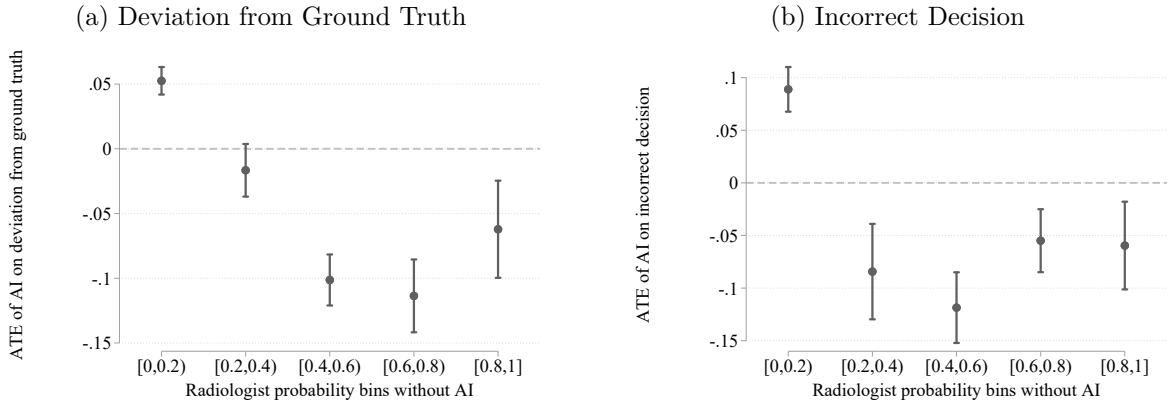
We next ask whether the information treatments affect radiologists' diagnostic performance. As a measure, we consider the absolute deviation of the radiologist's probability report from the binary ground truth, $Y_{irt} = \left|p_r\left(\omega_i = 1 \middle| s_{irt}\right) - \omega_i\right|$ where lower values implies better performance. Appendix C.4.2 contains results that use a continuous ground truth, which are

qualitatively similar. The middle panel of table 2 shows the average treatment effects on performance.

Our results indicate that while access to contextual information improves performance on average, AI assistance does not. We find that access to clinical history reduces the deviation from the ground truth by 4.8% ($p < 0.05$) of the control mean if we pool designs and by 6.7% in design 1 ($p < 0.05$). In contrast, the effects for AI are close to zero and not statistically significant. In light of the findings in the previous two sections — that the AI is more accurate than most radiologists and that radiologists move their assessments toward the AI—it is puzzling that the AI information does not improve accuracy.

This contradiction occurs because the average treatment effects mask significant heterogeneity in treatment effects. Our within-participant designs–designs 2 and 3–allow us to estimate conditional treatment effects given radiologists' predictions without AI assistance. Specifically, we partition cases based on the expert's signal into five equally spaced bins of $\pi_r\left(\omega_i = 1|s_{ir}^E\right)$. Figure 2 shows the conditional treatment effects of providing AI assistance on diagnostic performance. Panel (a) shows the deviation from the ground truth and panel (b) shows the probability of incorrect decision. We find that providing AI assistance in cases when the radiologist is uncertain improves performance on both metrics, whereas AI assistance is harmful when the radiologist is close to certain that the pathology is not present for a given case. Given an average reported probability of 25% or less, the vast majority of cases fall in the first bin yielding a small average treatment effect that masks important heterogeneity. The result that AI assistance can decrease performance rejects a model in which radiologists are Bayesians with correct beliefs.

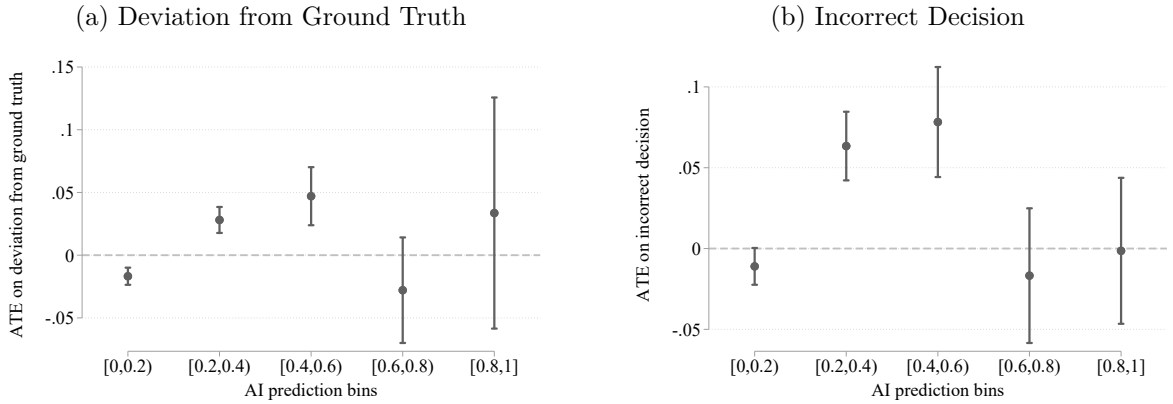
Figure 2: Conditional treatment effect given radiologist prediction



Note: Panel (a) shows the conditional average treatment effect of providing AI information on the absolute value of difference between the radiologist probability and the ground truth. Panel (b) shows analogous treatment effects on incorrect diagnosis, where a correct diagnosis is defined as the treatment recommendation matching the ground truth. Both these treatment effects are conditional on the ranges of the expert's prediction constructed using the no AI assistance. Standard errors are two-way clustered at the radiologist and patient-case level. The error bars depict 95% confidence intervals. Robustness to experimental design is in appendix C.4.1.

While AI assistance can help uncertain experts, we find that providing uncertain AI predictions reduces performance. As with the analysis of conditional treatment effects given expert predictions, we estimate conditional treatment effects given AI predictions by partitioning cases into five bins based on $\pi\left(\omega_i = 1|s_i^A\right)$. Figure 3 presents the estimates, pooling data from all three experimental designs. When the AI provides a confident prediction (e.g. either close to zero or close to one) performance is significantly improved. We see that in the lowest bins of AI signals, the deviation from the ground truth is reduced. In the second highest bin we also see a marked, though not statistically signficant, improvement in performance. However, in the middle range of signals, where the confidence of the AI is low, radiologists' diagnostic performance and probability of making a correct decision is lower when AI information is provided. This result reinforces the conclusion that radiologists err when updating their beliefs.

Figure 3: Conditional treatment effect given AI prediction



Note: Panel (a) shows the conditional average treatment effect of providing AI information on the absolute value of difference between the radiologist probability and the ground truth. Panel (b) shows analogous treatment effects on incorrect diagnosis, where a correct diagnosis is defined as the treatment recommendation matching the ground truth. Both these treatment effects are conditional on the ranges of AI prediction. Standard errors are two-way clustered at the radiologist and patient-case level. The error bars depict 95% confidence intervals. Robustness to experimental design is in appendix C.4.1 and C.4.2.

4.2.3 Treatment Effects on Time Per Case and Proxies of Effort

Finally, we turn our attention to the effects of AI assistance on time taken and the number of unique interaction (clicks) as a proxy for effort. One hypothesis is that AI assistance could economize on costly human effort without sacrificing overall performance by enabling quicker assessments. At the polar opposite, it is possible that humans take more time because they are provided with more information to process. Which of these effects dominate determines the effect on labor costs when humans use AI assistance, and therefore the optimality of delegating cases versus a collaborative setup.

Our results indicate that radiologists are slower when provided with AI assistance. The last panel of table 2 shows the treatment effects on time spent per case and clicks. These outcomes are measured at the case level. In the X-Ray Only treatment, radiologists spend about 2.6 minutes per case. Both AI and CH increase the time spend per case by a statistically significant amount of approximately 4%. The interaction effct $\gamma_{AI\times CH}$ is not significant for either of the two outcome variables. These effects suggest that decisions where both radiologists and the AI are involved come at a non trivial increase in time spent per case. This result further undercuts the potential benefits in performance from including humans assisted with AI predictions "in the loop."

4.3 Discussion

We find that AI assistance does not improve the performance of our participants on average, even though the AI predictions are more accurate than the majority of radiologists, and that radiologists respond to this assistance. However, the average treatment effects mask important heterogeneity – AI assistance improves performance for a large set of cases, but also decreases performance in many instances. These results point to biases in the use of AI predictions which we will further investigate in section 5.

We also find that humans have access to valuable contextual information, suggesting that full automation has its drawbacks. But, the biases above – especially in conjunction with our finding that radiologists take longer when given AI assistance – undercut this potential information advantage in a setup that involves AI assistance. Thus, the problem of how best to deploy AI assistance may be non-trivial and the optimal solution may involve selective automation and/or AI assistance. Section 6 analyzes this problem.

Appendix C.4 shows that the results are qualitatively robust to a variety of alternative analyses. The results are similar when we split the analysis by experimental design and when we only focus on the initial across-comparison (using design 1 and 2), although the latter leads to estimates that are imprecise. Alternative ground-truth measures – including a leave-one-out ground truth based on the assessments of our experimental participants and a continuous ground truth, which simply averages the assessments of the ground-truth labelers–also yield similar conclusions. The qualitative patterns of the treatment effects are unchanged if we calibrate each radiologists' assessments to the ground-truth before conducting the analysis. Finally, incentives for accuracy, which are cross-randomized in designs 1 and 3, do not have significant effects either.

5 Automation/Own-Information Bias/Neglect

An upshot of the results in section 4 is that our participants deviate from the baseline of a Bayesian with correct beliefs about the joint distribution of their own information and the AI signal. In this section, we model and estimate systematic deviations from this benchmark – which we will refer to as Bayesian[24] for short – and determine the implications of these deviations on utilizing human expertise and AI predictions.

[24]The omission of the qualifier "with correct beliefs" slightly abuses terminology because a possible explanation of the deviations we have documented is that our participants are Bayesians but update their beliefs using an incorrect model for the joint distribution of s^A, s^E and ω. We will entertain this possibility below.

5.1 A Model of Deviations from Bayesian Updating

The framework in section 2 shows that a key question is whether the odds-ratios

$$\frac{p_r\left(\omega_i = 1|s_i^A, s_{ir}^E\right)}{p_r\left(\omega_i = 0|s_i^A, s_{ir}^E\right)} \text{ and } \frac{\pi_r\left(\omega_i = 1|s_i^A, s_{ir}^E\right)}{\pi_r\left(\omega_i = 0|s_i^A, s_{ir}^E\right)}$$

differ from each other. Recall that Bayes' rule implies that

$$\ln \frac{\pi_r\left(\omega_i = 1|s_i^A, s_{ir}^E\right)}{\pi_r\left(\omega_i = 0|s_i^A, s_{ir}^E\right)} = \ln \frac{\pi_r\left(s_i^A|\omega_i = 1, s_{ir}^E\right)}{\pi_r\left(s_i^A|\omega_i = 0, s_{ir}^E\right)} + \ln \frac{\pi_r\left(\omega_i = 1|s_{ir}^E\right)}{\pi_r\left(\omega_i = 0|s_{ir}^E\right)}. \tag{4}$$

We now consider a set of models of belief updating to describe systematic deviations from this benchmark. In our model, the human correctly interprets their own signal when AI assistance is not available but errs when both s_i^A and s_{ir}^E are observed. As we will show below, AI assistance nonetheless unambigiously improves performance for a subset of parameters within this family.

The first class of biases that we consider, arises when the two terms on the right-hand side of equation (4) are incorrectly weighted. Following Grether (1980; 1992), we parametrize this type of error using the following parsimonious functional form:

$$\log \frac{p_r\left(\omega_i = 1|s_i^A, s_{ir}^E\right)}{p_r\left(\omega_i = 0|s_i^A, s_{ir}^E\right)} = b_r \log \frac{\pi_r\left(s_i^A|\omega_i = 1, s_{ir}^E\right)}{\pi_r\left(s_i^A|\omega_i = 0, s_{ir}^E\right)} + d_r \log \frac{\pi_r\left(\omega_i = 1|s_{ir}^E\right)}{\pi_r\left(\omega_i = 0|s_{ir}^E\right)}, \tag{5}$$

where $b_r, d_r \geq 0$. The Bayesian (with correct beliefs) is a special case with $b_r = d_r = 1$. While this linear form is restrictive, it has been useful for documenting several empirical regularities showing deviations from Bayesian updating like base-rate neglect and under inference (see Benjamin, 2019, for a review).

We will say that the human exhibits *automation bias* if $b_r > d_r$, and *automation neglect* if $b_r < d_r$. As a motivation for this nomenclature, observe that when $b_r > d_r$, the human over-weights the AI signal relative to their own. In the specific case when $d_r = 1$, the agent overshoots when updating the posterior odds relative to a Bayesian. Analogously, if $b_r < d_r$, then the human under-weights the AI signal relative to their own. We say that the human exhibits *own-information neglect* if $d_r < 1$ and *own-information bias* if $d_r > 1$, where the cutoff value of one is based on the Bayesian benchmark. Own-information biases are similar to base-rate biases, but apply to beliefs given the expert's signals instead of unconditional population rates (see Griffin and Tversky, 1992; Kahneman and Tversky, 1973).

A second class of deviations we consider will allow for models in which decision-makers do

not correctly account for the dependence between s_i^A and s_{ir}^E, which we call *signal dependence neglect*. For example, if humans act as-if s_i^A and s_{ir}^E are independent conditional on ω_i even if they are not, then their posterior beliefs can be written as

$$\log \frac{p_r\left(\omega_i = 1 | s_i^A, s_{ir}^E\right)}{p_r\left(\omega_i = 0 | s_i^A, s_{ir}^E\right)} = b_r \log \frac{\pi_r\left(s_i^A | \omega_i = 1\right)}{\pi_r\left(s_i^A | \omega_i = 0\right)} + d_r \log \frac{\pi_r\left(\omega_i = 1 | s_{ir}^E\right)}{\pi_r\left(\omega_i = 0 | s_{ir}^E\right)}, \tag{6}$$

where b_r and d_r are allowed to differ from 1 as above. In the case when the signals are jointly multivariate normal and $b_r = d_r = 1$, signal dependence neglect yields correlation neglect as defined in (Enke and Zimmermann, 2019).[25] More generally, we will consider models that vary the conditioning set in the first term on the right-hand side and the dimension of s_i^A in the first term on the right-hand side. The specific examples are motivated and discussed further in section 5.3 below.

We intend the functional form above as descriptions of humans' updating rules and will remain agnostic about underlying mechanisms and micro-foundations. In particular, we remain silent on whether our participants are Bayesians that are utilizing the incorrect joint distribution of $\left(\omega_i, s_i^A, s_{ir}^E\right)$ when updating their beliefs or if they are non-Bayesians. The former type of model, known as a quasi-Bayesian model,[26] can generate automation bias or neglect as well as correlation biases.[27] An implicit assumption in our model, and likely other micro-foundations for the functional forms above as well, is that the signal acquired by the human is invariant to the provision of AI assistance. Whether additional training, or experience with the AI can correct deviations from the benchmark model is therefore something that we leave for future work.

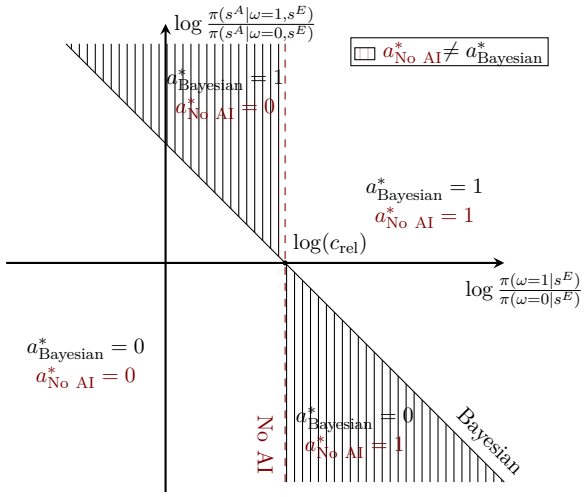
Nonetheless, the models above will prove useful for our purposes. From a theoretical perspective, they will help outline the types of deviations that potentially decrease decision quality. From an empirical perspective, they help understand the drivers of the treatment effects documented earlier and turn out to be a good approximation to the data from the experiment.

[25] If s_i^A and s_{ir}^E are unidimensional with $\left(s_i^A, s_{ir}^E\right) \sim N\left(0, \Sigma_r\right)$, then the covariance matrix Σ_r is a sufficient statistic for the posterior probability that $\omega_i = 1$ given the signals if $\omega_i = 1\left\{s_i^A + s_{ir}^E \geq \varepsilon_i\right\}$ and ε_i is independent of $\left(s_i^A, s_{ir}^E\right)$.

[26] See Rabin (2013) for a definition and Barberis et al. (1998); Rabin (2002); Rabin and Vayanos (2010) for examples.

[27] To see this, assume that $p_r\left(s_i^A | \omega_i, s_{ir}^E\right) = \pi\left(s_i^A | \omega_i, s_{ir}^E\right)^b$ and $p_r\left(s_{ir}^E, \omega_i\right) = \pi_r\left(s_i^E, \omega_i\right)$ to generate the functional form in equation (5) for any b_r as long as $d_r = 1$. The derivation of equation (6) is similar. In contrast to automation bias/neglect and correlation biases, own-information bias/neglect cannot be derived in a quasi-Bayesian model because we assume that $p_r\left(\omega_i | s_{ir}^E\right) = \pi_r\left(\omega_i | s_{ir}^E\right)$.

Figure 4: Comparing decisions with and without AI assistance – Bayesian with correct beliefs



Note: The figure shows the decision criterion of a Bayesian with and without AI assistance and where their decisions agree and disagree. Shaded regions show the regions in which AI improves or worsens decision making.

5.2 Implications for Human-AI Collaboration

We now show that the types of deviations described above have implications for when AI assistance unambiguously improves human performance. The results will also illustrate the utility of the simple functional forms in equations (5) and (6). This subsection drops the i and r indices for simplicity of notation.

It is useful to start by considering the decisions with and without AI assistance for a Bayesian decision maker. Figure 4 illustrates the realizations of s^A for which the optimal decision with AI assistance differs from the the decision without AI assistance for a fixed c_{rel}. The horizontal and vertical axes respectively represent $\log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ and $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$. As shown by the vertical dashed line, the decision-maker would take the action 1 if and only if $\log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ exceeds $\log c_{rel}$. The solid line represents the analogous boundary for a Bayesian who has access to AI assistance. Observe that the decisions a Bayesian makes as a function of the signals s^A and s^E cannot be improved without additional information. Thus, a Bayesian with correct beliefs and access to both signals improves upon the no AI action in the vertically shaded region.

Now consider humans that may deviate from this benchmark model. A human who takes a given action without AI assistance $a^*_{\text{No AI}} = a^*\left(s^E; p_r\right)$, but a different action with AI assistance (so that $a^*_{\text{No AI}} \neq a^*_{\text{AI}}$) makes a worse decision if $a^*_{\text{AI}} = a^*\left(s^A, s^E; p_r\right)$ disagrees with

the Bayesian's decision $a^*_{\text{Bayesian}} = a^*\left(s^A, s^E; \pi_r\right)$ with AI assistance. This follows because, in the binary action setup, only one of the decisions can agree with the Bayesian decision. In all other cases, the human's decision is weakly improved for the signal realization $\left(s^A, s^E\right)$. In other words, a human whose decision changes upon receiving the AI signal s^A is better off with AI assistance only if the change agrees with the Bayesian decision. The human is unambiguously better off if this property holds for all signals.

Our first result states that a human who deviates from the benchmark model because they only exhibits automation neglect is unambiguously better off with AI assistance.

Proposition 1. *Suppose that the human's posterior is described by equation (5).*

(i) If the human only exhibits automation neglect $(b < 1, d = 1)$*, then for all pairs of signal realizations* $\left(s^A, s^E\right)$*, and any* c_{rel}*, the human attains weakly higher expected payoff* $V(s)$ *with AI assistance.*

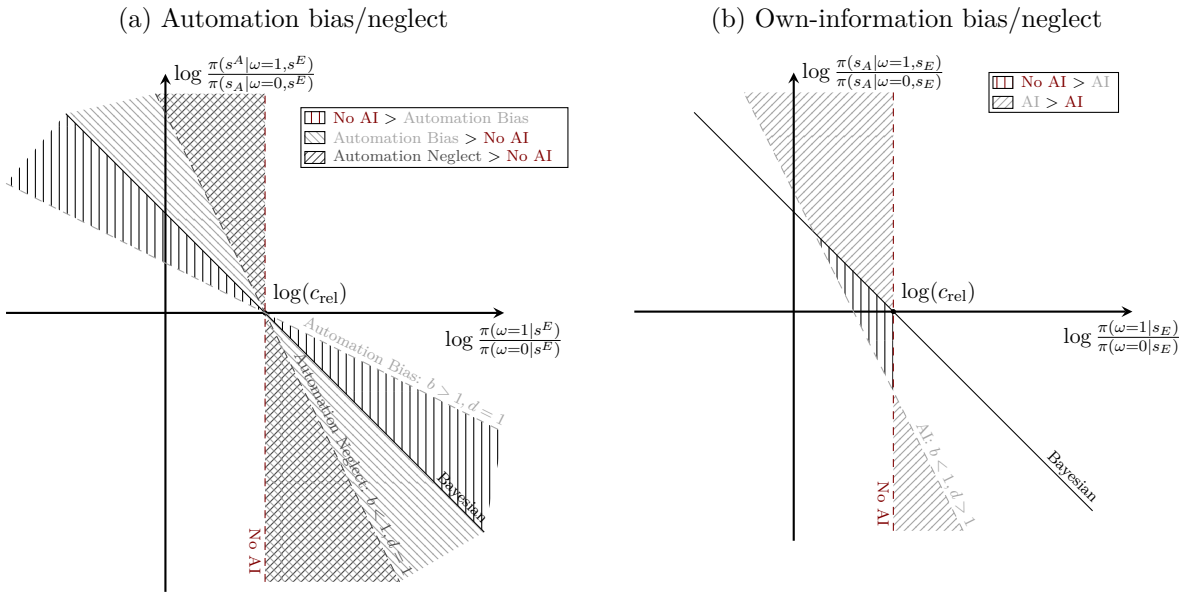
(ii) In all other cases $(b > 1$ *or* $d \neq 1)$*, for any* c_{rel}*, there exist log-likelihood ratios* $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$ *and* $\log \frac{\pi\left(\omega=1|s^E\right)}{\pi(\omega=0|s^E)}$ *such that the human attains lower expected payoff* $V(s)$ *with AI assistance.*

See appendix A for the proof.

Figure 5a illustrates the case without own-information bias or neglect by setting $d = 1$ but allows for either automation bias or neglect by setting $b \neq 1$. The two dashed lines represent cutoffs analogous to those in figure 4 for humans with automation bias and automation neglect. Although a human that only exhibits automation neglect under-responds to the AI information, their beliefs move towards a Bayesian decision-makers but does not overshoot it. Whenever her decision changes, it agrees with the Bayesian's. In contrast, if the human exhibits automation bias, they errs for moderately informative AI signals with intermediate values of $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$ because they over-reacts. At high enough values of this log-likelihood ratio, both the Bayesian and the human exhibiting automation bias would take the same action.

In the case when there is either own-information bias or neglect $(d \neq 1)$, it is always possible for the payoff with AI to be lower than the payoff without AI assistance. In our model, this occurs because providing the AI signal generates a bias in how the expert uses their own information. Figure 5b illustrates the result (for $d > 1$ and $b < 1$). Relative to the analogous line in figure 5a, the intercept of the dashed line is not equal to $\log c_{rel}$. The vertically striped triangle depicts combinations of the log-likelihood ratios above in which the biased expert achieves a higher payoff without AI. In these cases, the expert takes the correct action without AI assistance but incorrectly uses their own information with AI assistance, resulting in an

Figure 5: Comparison with Bayesian decisions



Note: This figure shows the where the decisions of an expert as a function of the signals disagree with the Bayesian in cases with and without AI assistance. Panel (a) shows automation bias and neglect (absent own-information bias and neglect). Panel (b) shows a decision maker who exhibits both automation neglect and own-information bias.

error. Hence, in the presence of own-information bias or neglect, AI assistance can result in worse decision for some signals.

We next consider a decision-maker with a different type of bias, namely, one in which the decision-maker exhibits signal dependence neglect. Perhaps not surprisingly, our next result shows that this type of bias on its own can result in worse decisions with AI assistance:

Proposition 2. *Suppose that the human exhibits signal dependence neglect so that the posterior belief is described by equation (6). For any value of* $b > 0$, $d > 0$, *and* $c_{rel} > 0$, *the there exist log-likelihood ratios* $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$ *and* $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)}$ *such that the human attains lower expected payoff* $V(s)$ *with AI assistance.*

Compared to the models of only automation/own-information bias/neglect, signal dependence neglect adds another dimension of potential mistakes to those illustrated in the figures above. The result bears resemblance to those in Enke and Zimmermann (2019), which showed that in a multivariate normal model with positively correlated signals, correlation neglect results in over-reaction to signals and verified this hypothesis in lab experiments. Proposition 2 differs in that it allows for general signal distributions and for signal dependence neglect to co-exist with automation and own-information bias/neglect. This extension is essential for a naturalistic environment like ours because the experimenter does not have full control over

the signal structure. The general signal structure makes it difficult to characterize mistakes in terms of over or under-updating, unlike in the case of a multivariate normal model. Even when $b = d = 1$ so that automation or own-information bias/neglect are not relevant, an examination of equations (4) and (5) reveals that whether or not a decision-maker exhibits under- or over-updating depends on the difference between $\log \frac{\pi\left(s^A|\omega=1,s^E\right)}{\pi(s^A|\omega=0,s^E)}$ and $\log \frac{\pi\left(s^A|\omega=1\right)}{\pi(s^A|\omega=0)}$.

The propositions above have important implications for the design of human-AI collaboration, which we consider in section 6. Specifically, we study an AI designer who only has access to the AI signal s^A and must decide on one of the three modes of delegation – utilize only the AI prediction, delegate the case to the human, or provide AI assistance to a human expert. The results show that other than in the case when automation neglect is the only relevant bias, the designer must learn the types of biases as well as the distribution of $\pi\left(s^A, s^E|\omega\right)$ to determine which delegation modality yields the best decision.

5.3 Estimating Deviations from Bayesian Updating

We now turn to an empirical implementation of the model above. The analysis in this section will be based on designs 2 and 3 because they allow us to observe the same participant make decisions under all information-conditions on a given case. Start by considering an empirical analog to equation (5):

$$\log \frac{p_r\left(\omega_i = 1|s^A_{ir}, s^E_{ir}\right)}{p_r\left(\omega_i = 0|s^A_{ir}, s^E_{ir}\right)} = a + b \cdot \log \frac{\pi_r\left(s^A_{ir}|\omega_i = 1, s^E_{ir}\right)}{\pi_r\left(s^A_{ir}|\omega_i = 0, s^E_{ir}\right)} + d \cdot \log \frac{\pi_r\left(\omega_i = 1|s^E_{ir}\right)}{\pi_r\left(\omega_i = 0|s^E_{ir}\right)} + \varepsilon_{ir}, \quad (7)$$

where we have omitted heterogeneity across radiologists in b_r and d_r. Appendix C.5.6 discusses radiologist heterogeneity in our estimates. Two of the terms in this equation are directly elicited in the experiment: the probability in the second term on the right-hand side, $\pi_r(\omega_i = 1|s^E_{ir})$, is equal to the radiologists' assessment in the treatment arm where the subjects read cases without AI assistance and the term $p_r(\omega = 1|s^A_{ir}, s^E_{ir})$ in the dependent variable is the assessment in the treatment arm with AI. The "update term," given by $\log \frac{\pi_r\left(s^A_{ir}|\omega_i=1,s^E_{ir}\right)}{\pi_r\left(s^A_{ir}|\omega_i=0,s^E_{ir}\right)}$ will be estimated and substituted into the equation above.

There are three challenges in estimating the update term. The first challenge is that it is a ratio of conditional densities. We address this issue by rewriting it using Bayes' rule as

follows:

$$\log \frac{\pi_r\left(s_{ir}^A|\omega_i=1,s_{ir}^E\right)}{\pi_r\left(s_{ir}^A|\omega_i=0,s_{ir}^E\right)}=\log \frac{\pi_r\left(\omega_i=1|s_{ir}^A,s_{ir}^E\right)}{\pi_r\left(\omega_i=0|s_{ir}^A,s_{ir}^E\right)}-\log \frac{\pi_r\left(\omega_i=1|s_{ir}^E\right)}{\pi_r\left(\omega_i=0|s_{ir}^E\right)}.$$

If s_{ir}^E can be constructed or controlled for, then we can estimate the first term on the right-hand side using data on ω_i and s_{ir}^A via a binary response model. This estimation is possible because the signal from the AI that the humans receive is isomorphic to the vector of predicted probabilities for the various diseases observed. The second term in this equation has been elicited.

This brings us to the second challenge – constructing s_{ir}^E when estimating $\pi_r\left(\omega_i=1|s_{ir}^A,s_{ir}^E\right)$ because we do not observe it directly. Let c_i be the patient case associated with case-pathology i and $I\left(c_i\right)$ be the set of case-pathologies associated with case c_i. If s_{ir}^E is unidimensional and $\pi_r\left(\omega_i|\, s_{ir}^E\right)$ is monotonic in s_{ir}^E, then $\pi_r\left(\omega_i|\, s_{ir}^E\right)$ is a valid control variable. However, we want to allow for the possibility that the radiologist evaluates a case holistically and uses signals across pathologies. Under the assumption that $s_i^A \perp s_{ir}^E \big|\, \omega_i, \left(\pi_r\left(\omega_{i'}|\, s_{i'r}^E\right)\right)_{i'\in I(c_i)}$, the vector of probabilities for all pathologies reported by r for case i, denoted $\left(\pi_r\left(\omega_{i'}|\, s_{i'r}^E\right)\right)_{i'\in I(c_i)}$, is a valid control variable. In this multi-dimensional case, a sufficient condition is that s_{ir}^E is invertible in $\left(\pi_r\left(\omega_{i'}|\, s_{i'r}^E\right)\right)_{i'\in I(c_i)}$. Our empirical specifications will therefore employ multivariate proxy controls for s_{ir}^E using elicited probability assessments for multiple pathologies.

To allow for flexible interactions between s_i^A and s_{ir}^E while avoiding over-fitting, we estimate $\pi_r\left(\omega_i=1|s_{ir}^A,s_{ir}^E\right)$ using a pathology-specific random forest that predicts ω_i using the vector of predicted probabilities for all pathologies for case c_i reported by radiologist r without AI assistance, the vector of predicted probabilities for case c_i the AI algorithm produces, and summaries of the patient clinical history when made available to the radiologist, and participant-specific fixed-effects. The hyper-parameters of the random forest are chosen by grouped k-fold cross-validation, where we ensure that each patient case appears in only one fold to avoid overfitting to the patient case. Further details of the training procedure are described in appendix C.5.2.

The third challenge is the potential for measurement error, particularly in the form that radiologists' signal s_{ir}^E when elicited without AI might differ from their signal when given AI assistance. Classical measurement error arising from this source would lead to attenuation bias in the coefficient estimates. To address this issue, we will construct instruments s_{ir}^E using the reported probabilities of the other radiologists in our experiment.

With these solutions in hand, we would like to assess whether humans exhibit signal depen-

dence neglect when updating beliefs. As prefaced earlier, although the humans' and AI's signals are not conditionally independent given the ground truth, humans may act as-if they are. We will therefore estimate and select between models that vary the set of signals conditioned on in the update term. For example, in the case when radiologists behave as if s_i^A and s_{ir}^E are independent conditional on ω_i, the update term drops the conditioning on s_{ir}^E. We can vary the pathologies across the set of models considered when constructing $I(c_i)$.[28] Including these models in our selection approach will also tell us whether radiologists' assessments are separable across different pathologies.

The correct model of behavior satisfies the conditional moment restriction $E\left[\varepsilon_{irt} \middle| s_{i,-r}^E, s_i^A\right] = 0$, where $s_{i,-r}^E$ collects the signals of the radiologists other than r in our experiment. For estimation, we utilize unconditional moment restrictions based on functions of $s_{i,-r}^E$ and s_i^A that closely mimic the terms in equation (7). Our instruments include $\log \frac{\pi\left(\omega_i=1|s_i^A\right)}{\pi\left(\omega_i=0|s_i^A\right)}$, and leave-one-out averages of $\log \frac{\pi\left(\omega_i=1|s_i^A, s_{ir'}^E\right)}{\pi\left(\omega_i=0|s_i^A, s_{ir'}^E\right)}$ for radiologists other than r that use various proxies for $s_{ir'}^E$ based on different sets of pathologies $I(c_i)$ that are conditioned on.[29] Empirical analogs of the resulting moment conditions are used to estimate the model using GMM.

We will employ the model-selection procedure proposed in Andrews and Lu (2001) to select between non-nested models. This method utilizes the J-statistic of the GMM objective function, which is adjusted for the number of moments and parameters that are included in the model. The selection criterion, MMSC-BIC, penalizes models that reject a greater number of moment restrictions.

5.4 Results

Table 3 presents estimates from six models, which is a subset of the models we consider.[30] The first three models do not utilize clinical history when made available to the participants to construct the proxy for s_{ir}^E whereas the last three models do. We consider three models in each of these two sets by varying $I(c_i)$. The first corresponds to the case when the signal s_i^A only includes the focal pathology and is conditionally independent of s_{ir}^E given ω_i. The

[28] The conditionally independent case corresponds to the extreme case in which $I(c_i) = \emptyset$, whereas the Bayesian model includes all pathologies.

[29] Specifically, we construct fourteen instruments. The first is a constant and the second is the average of $\log \frac{\pi\left(\omega_i=1|s_{ir'}^E\right)}{\pi\left(\omega_i=0|s_{ir'}^E\right)}$ for all $r' \neq r$. The remaining twelve construct the average of $\log \frac{\pi(\omega_i=1|s_{ir'})}{\pi(\omega_i=0|s_{ir'})}$ for all $r' \neq r$ by varying the conditioning variables s_{ir}. The different sets of conditioning variables in s_{ir} are presented in the second panel of appendix table C.21. These sets are used because they are the relevant terms in at least one of the models that we consider in the testing procedure.

[30] See table C.20 in the appendix for the results from all models. Results from all pathologies with AI are qualitatively similar (see table C.21).

second allows for dependence between s_i^A and s_{ir}^E given ω_i, but only through $\pi\left(\omega_i|\, s_{ir}^E\right)$ for the focal pathology and potentially clinical history. The third utilizes the correct update term constructed using all available AI predictions and the vector of probabilities $\pi\left(\omega_i|\, s_{ir}^E\right)$ for all pathologies. Setting $b = d = 1$ and the constant to 0 in this last case corresponds to Bayesian updating with correct beliefs.

The results from this exercise point to two types of errors in radiologists' use of AI signals. The first type of error is that radiologists neglect signal dependence even though AI predictions and radiologists' signals are highly correlated after conditioning on the ground truth (see appendix table C.18). This conclusion follows because we select the model in column 1, which uses an update term based on equation (6) and therefore does not control for s^E. Another implication of the selected model is that radiologists do not incorporate information across different pathologies since only the focal pathology is relevant. This result validates our previous analysis that analyzes each pathology separately.

Second, across different models, including all models that are not selected, we estimate that radiologists exhibit automation neglect. However, radiologists do not exhibit substantial own-information bias or neglect since the estimated value of d is close to 1 in all specifications. Accounting for measurement error in estimating this model is important, as versions that do not do so estimate d significantly less than 1 (appendix C.5.3) .

Connecting these observations back to our theoretical discussion in section 2, the parameters are such that radiologists will not benefit unambiguously from access to the AI signal, primarily because of signal dependence neglect.

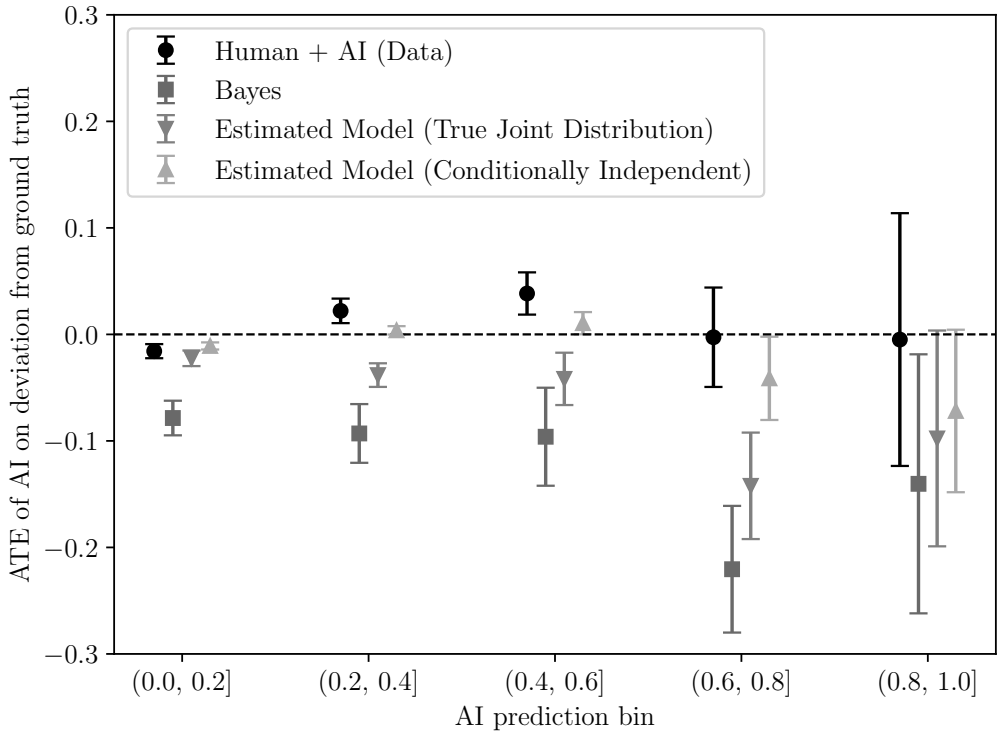
Table 3: Model Selection: Top Level with AI – IVGMM

	(1)	(2)	(3)	(4)	(5)	(6)
Automation Bias (b)	0.29	0.34	0.35	0.19	0.27	0.35
	(0.02)	(0.03)	(0.03)	(0.02)	(0.02)	(0.03)
Own Info Bias (d)	1.10	1.08	1.05	1.07	1.06	1.05
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Constant	0.40	0.40	0.37	0.31	0.33	0.36
	(0.03)	(0.04)	(0.04)	(0.03)	(0.03)	(0.04)
Focal s_A	✓	✓	✓	✓	✓	✓
Other s_A			✓			✓
Focal s_E		✓	✓		✓	✓
Other s_E			✓			✓
Clinical History s_E				✓	✓	✓
J-Statistic	18.23	16.86	0.17	4.3	10.7	0.09
MMSC-BIC	-23.96	-21.12	-8.27	-12.57	-10.39	-8.35
Number of Moments	13	12	5	7	8	5
Observations	11420	11420	11420	11420	11420	11420
R-Squared	0.50	0.49	0.52	0.46	0.48	0.52

Note: This table summarizes estimates of b and d for different specifications of the update term. The models differ by whether the update term conditions on the signal s_E of the pathology at hand, the AI's and the radiologist's signal of other pathologies as well as the information that is provided in the clinical history of a patient. Each model is estimated via GMM. The reported MMSC-BIC statistic adjusts the J-statistic for the number of included parameters, awarding bonus terms for models with fewer parameters (see Andrews and Lu (2001) for details). This table presents a subset of the models that we select from. The full set of models included in the selection procedure are presented in table C.20. The update term is estimated via random forest as described in appendix section C.5.2. Standard errors are clustered at the radiologist level.

These deviations also explain the heterogeneous conditional average treatment effects documented in section 4. Figure 6 shows the treatment effect of AI on the deviation from ground truth of our radiologists along with three different models under which we compute the same treatment effects: the Bayesian benchmark, the model in equation (5) where radiologists use the correct updating term, and the selected model from table 3. As expected, the Bayesian would benefit significantly from gaining access to the AI signal. In fact, as indicated by the large reductions in the deviation from the ground truth, there is significant potential value in combining the human and the AI signal. Imposing the model in equation (5) with the correct updating term – column (6) – reduces these improvements and moves the implied treatment effects closer to the data. However, we see that throughout the entire signal range of s_A, such a decision-maker would still unambiguously benefit from gaining access to the AI. Only when we impose the selected update term, under which radiologists do not account for the dependence of signals, does the model generate the observed negative treatment effect

Figure 6: Model Treatment Effects



Note: This graph shows the observed conditional treatment effects of providing radiologists access to AI (Human+AI) and compares those to three different model-implied treatment effects: giving AI access to a Bayesian decision maker, giving AI access to a decision maker who acts according to the empirical version of equation 5, both under the correct updating term and when the decision maker treats the AI signal as conditionally independent.

conditional on intermediate AI predictions as in our treatment effect analysis. We see that such a model closes even more of the gap between the Bayesian treatment effects and the observed treatment effects.

It is significant that the model presented above is able to replicate the pattern of conditional treatment effects even though it is restrictive – it imposes linearity in log-odds and does not allow for heterogeneity in b and d across radiologists. Appendix C.5.5 investigates the linearity restriction in equation (7). We estimate a boosted tree that predicts the dependent variable in the equation using the same two independent variables in the equation. We find that the empirical analog to the relevant boundary of the decision-regions depicted in figure 5 is well-approximated using a linear model. Appendix C.5.6 investigates heterogeneity – it presents estimates that allow for b and d to vary with radiologists. We find that the estimated distributions of b_r and d_r are centered close to the point estimates above, with most of the estimated distributions of b_r and d_r in the range $[0.1, 0.4]$ and $[1.0, 1.2]$ respectively.

Thus, consistent with the conclusions above, we find evidence for automation neglect, but no meaningful own-information bias or neglect. These results suggest that the estimated model represents a good approximation to the experimental data.

Taken together, the results indicate that while there are large potential gains from combining radiologists' assessments with AI predictions, biases in radiologists' use of AI assistance undercuts these gains. We find that radiologists exhibit both automation-neglect and signal dependence neglect. These mistakes lead to no better diagnostic performance when provided with AI assistance. One approach for addressing this issue is to provide radiologists with better training, an avenue we leave for future research. A second approach is to design better collaborative systems that are built on predicting the types of cases in which human experts outperform AI predictions and vice-versa, and the types of cases in which AI-assisted humans do best, an issue we turn to next.

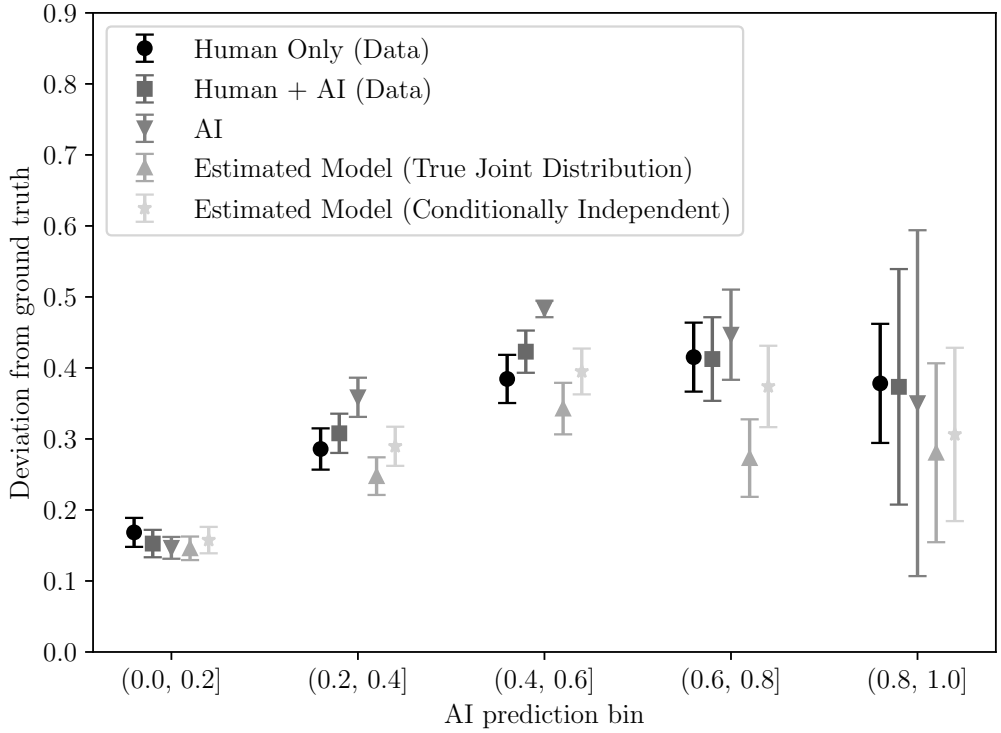
6 Designing Human-AI Collaboration

We now consider the design of collaborative systems between AI and radiologists. The designs we consider restrict attention to systems that use the AI signal to delegate a case to one of three modalities – radiologists alone, the AI alone, or the radiologist with access to the AI. As a warm-up for this exercise, it is useful to examine the predictive performance of the different modalities, conditional on s^A (see figure 7). Recall from the conditional treatment effect analysis, radiologists' assessments improve with AI in the lowest and the two highest bins of AI signals. Figure 7 also shows that in two out of three ranges in which the AI improves radiologists' decision-making that its own performance is even better than the performance of the radiologists with AI.

However, this figure misses the differences in the human time costs across modalities. Our analysis of the estimated treatment effects shows that radiologists take more time when provided with AI predictions. Moreover, using AI predictions alone saves on costly human effort. This points to the conclusion that in most cases where AI improves decision-making, one is at least as well off by relying exclusively on AI prediction. But using only AI or never assisting humans with AI predictions may come at the cost of performance in cases when using the human alone is superior.

Motivated by these observations, we now examine if there is a trade-off between the marginal costs of human effort and diagnostic performance.

Figure 7: Model Deviation from Ground Truth



Note: This figure shows the performance of the different modalities that we consider for the optimal collaborative system. Cases are either decided by only the radiologist, only the AI, or the radiologist with access to the AI. The performance measures for Human Only and Human + AI are constructed from our treatment effect analysis.

6.1 Computing the Trade-off Between Decision Loss and Costs of Human Effort

Consider a policy $\tau(\cdot)$ that chooses between full automation (AI), radiologist with access to AI ($E + AI$), or radiologist without access to AI (E), as a function of the AI signal s_i^A. The optimal policy which minimizes the sum of the expected decision-loss (costs of false positives and false negatives) and the monetized time cost of using human radiologists solves:

$$\tau^*(s_i^A) = \arg \min_{\tau \in \{E, E+AI, AI\}} -m \cdot V_{i\tau(s_i^A)} + w \cdot C_{i\tau(s_i^A)}. \tag{8}$$

The first term contains the expected decision-loss from a modality given by $V_{i\tau(s_i^A)} = E\left[V_{ir\tau\left(s_i^A\right)} \middle| s_i^A\right]$, which is the expected diagnostic quality averaging over radiologists and cases given the modality, preferences for false positives and false negatives, and the AI signal. Preferences for false positives and false negatives are allowed to vary by pathology but are the same across modalities. We estimate these preferences using data on the binary treatment recommendations of the radiologists in our experiment, given their probability assessments. Details of this preference estimation step are available in appendix C.6. The average cost of false negatives across pathologies is three times the cost of a false positive. Since we do not know the dollar cost of a false negative (m), we will present results for varying values of m.

The second term in the objective function contains $C_{i\tau(s_i^A)} = \left[C_{ir\tau\left(s_i^A\right)} \middle| s_i^A\right]$, which is the expected radiologist time for a given modality. If the case is fully automated, this time cost is zero and otherwise those time costs are based on our experimental estimates, which show that radiologists spend more time on cases when presented with AI predictions. For the interpretation of the magnitudes, we translate both radiologists' expected time cost and the expected decision loss into dollars. For the costs of radiologist time, we set $w = \$4$ per minute based on a payment of \$10 per case and an average time per read of 2.5 minutes.

We solve this problem by training a classification tree that assigns the case to a modality conditional on the AI signal $s_{A,i}$. We use 55% of cases as training data for the decision tree, 20% of cases for validation, and 25% in the test set. For this exercise, we focus on two top-level pathologies, cardiomediastinal abnormality, and airspace opacity. Finally, to quantify the drawbacks of biased humans, we will contrast the solution from such a system when the decisions under $E + AI$ are as observed in the data with the system that uses the Bayesian assessments for $E + AI$.

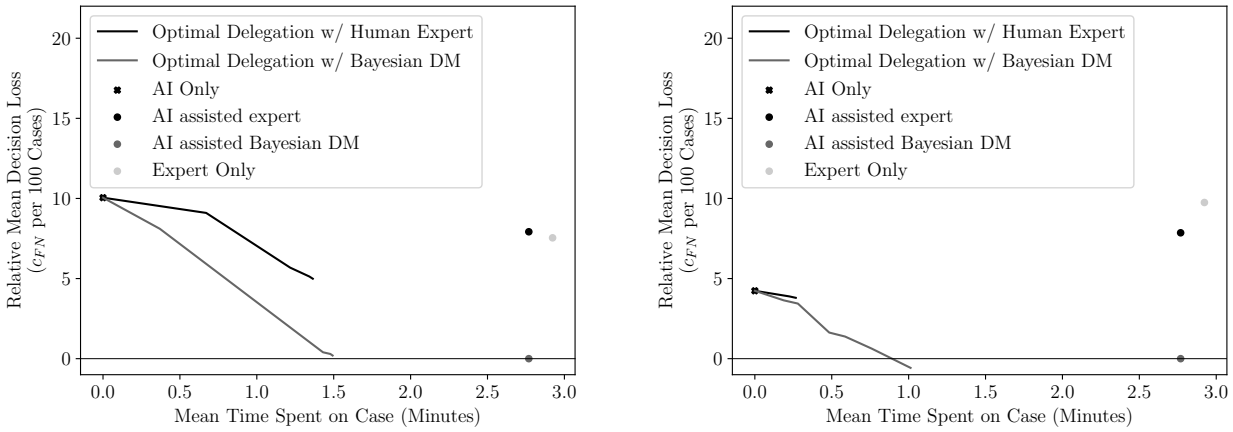
An important restriction in this formulation of the delegation problem is that the signal received by the expert is exogeneous and does not respond to the availability of AI. This assumption rules out models of endogenous information acquisition, for example, rational

inattention as in Sims (2003). A potentially interesting aspect is whether a designer can leverage such endogenous responses by designing an information revelation policy that induces effort. We leave such extensions that leverage insights from information design (Kamenica and Gentzkow, 2011; Bergemann and Morris, 2019) to future work, but measure the total amount of time taken by the experts in our experiment with and without AI assistance.

6.2 Results

Figure 8 shows a possibilities frontier for the trade-off between diagnostic quality against decision time for the two top-level pathologies with AI, calculated by varying m. It also shows the corner solutions where all cases are assigned to each of the modalities. The results suggest that there are potential gains from the optimal delegation of cases. An unassisted radiologist takes 2.7 minutes per case, or about $11, and incurs a relative decision loss of approximately seven for both of the top level pathologies with AI assistance. By moving to a solution on the frontier with a human expert – one whose assessments are as in our experiment – the designer can reduce both decision loss and the time taken per case by delegating many cases to the AI. The frontier with a Bayesian expert, as expected, is much more favorable. Moreover, the frontier for both a Human and a Bayesian expert coincide when the dollar cost of incorrect decisions (m) is small because the vast majority of cases are assigned to the AI.

Figure 8: Loss-Time Frontier



(a) Airspace Opacity

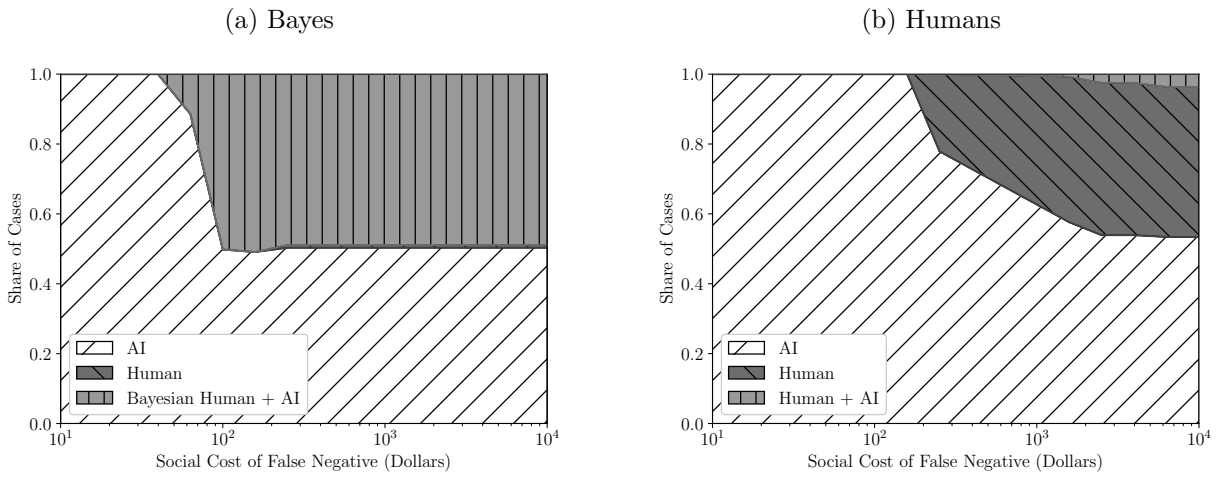
(b) Cardiomediastinal Abnormality

Note: This graph shows how radiologists and the AI perform relative to delegation systems on the frontier of the cost of human time versus decision loss. Panel (a) focuses on airspace opacity. Panel (b) focuses on cardiomediastinal abnormality.

Next, we investigate the proportion of decisions assigned to the three modalities as we vary

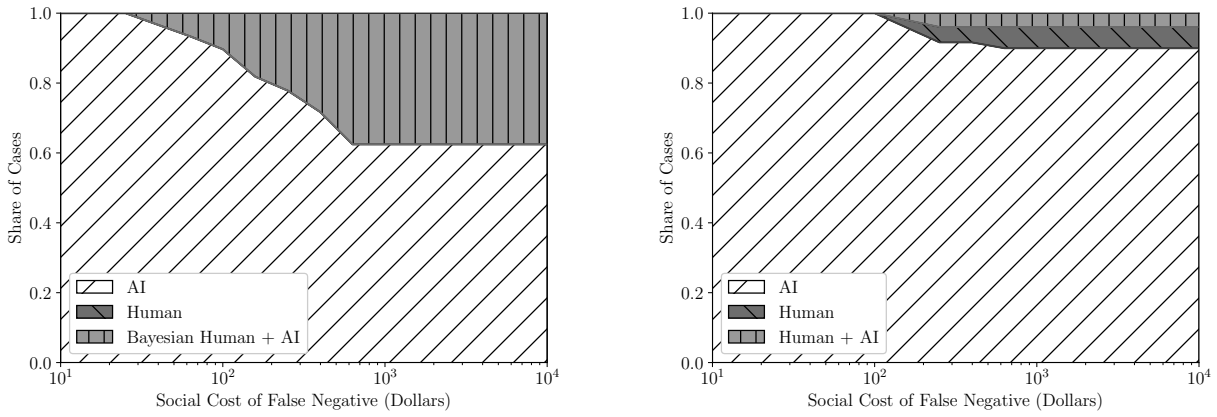
m (figures 9 and 10). Both for the case where radiologists are Bayesian and the observed behavior in our experiment, we find that the AI decides almost all cases if the cost of a false negative – a missed diagnosis – is less than \$100 per case. If radiologists acted as Bayesians with correct beliefs, the share of cases that involve human-AI collaboration rises markedly above a cost of \$100, but even for costs as high as \$10,000, this share does not exceed 50% for airspace opacity and does not exceed 40% for cardiomediastinal abnormality. Moreover, under Bayesian decision-making, the share of cases where only the radiologist decides is negligible for both pathologies and the only reason for using an unassisted radiologist is to save on time costs. When we conduct the same exercise and use the observed behavior of radiologists, we find that humans are involved in approximately 50% of cases for airspace opacity if the cost of a false positive is sufficiently large. For cardiomediastinal abnormality, the share of cases with humans involved only reaches 10% for very large costs of false positives. Moreover, the majority of cases where a radiologist is involved have the radiologist make decisions without AI for both pathologies. A more complete assessment of the optimal combination of human and machine decisions, therefore, confirms the intuition from above that cases are either decided by the radiologist or the AI but not by both of them together.

Figure 9: Airspace Opacity Modality Shares



Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for airspace opacity. Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment.

Figure 10: Cardiomediastinal Abnormality Modality Shares



(a) Bayes

(b) Humans

Note: The graphs show the share of cases decided by each modality (humans, AI, humans+AI) conditional on the cost of a false negative in dollars, denoted m in the text, for cardiomediastinal abnormality Panel (a) focuses on a Bayesian decision maker. Panel (b) focuses on a human decision-maker with decisions and time-taken as in our experiment.

6.3 Caveats

There are several caveats to the analysis here. The first is that any collaborative system may change radiologists' expectations about the difficulty of cases and adjust strategically to those changing expectations. For example, if radiologists are only handed the most difficult cases, they may exert more effort. We leave such strategic adjustments to a collaborative system for future work. We also have not considered alternative mechanisms to elicit radiologists' signals that may be less prone to errors. For instance, one may want to ask the radiologist to submit their assessment before seeing the AI's assessment and then combine both assessments optimally ex-post. Finally, the solutions presented here treat pathologies as separable. This case is relevant if there is a single pathology of interest.

7 Conclusion

AI is predicted to profoundly reshape the nature of work (see Felten et al., 2023). Humans are likely to use AI as decision aids for many tasks not only in the long run but also in the medium run for tasks that will ultimately be fully automated. A central question is therefore how humans use AI tools and how tasks should be assigned. To understand the benefits and pitfalls of human-machine collaboration, we conduct an experiment in which radiologists are provided with AI assistance. Besides serving as an iconic example of a

highly-skilled profession that is being transformed by AI, radiology also represents a large class of professionals whose main job is a high-stakes classification task. Since we can simulate radiologists' normal workflow, this is an ideal setting for conducting such an experiment.

While the potential benefits of deploying AI assistance are large in our setting, biases in humans' use of AI assistance eliminate these potential gains. Even though the AI tool in our experiment performs better than two thirds of radiologists, optimally combining their predictions could substantially improve performance. Yet, we find that giving radiologists access to AI predictions does not, on average, lead to higher performance. This average treatment effect, however, masks systematic heterogeneity: providing AI does improve radiologists' predictions and decisions for cases where the AI is certain, but not when uncertain. This latter result – that prediction quality can be reduced for some range of AI signals – rejects Bayesian updating with correct beliefs. We also describe systematic errors in belief updating – namely radiologists exhibit automation neglect (e.g. radiologists weight the AI prediction relative to their own) and treat the AI prediction and their own signals as independent even though they are not. Moreover, radiologists take significantly more time to make a decision when AI information is provided.

Together, these results have important implications for how to design the collaboration between humans and machines. Increased time costs and sub-optimal use of AI information both work against having radiologists make decisions with AI assistance. In fact, an optimal delegation policy that utilizes heterogeneity in treatment effects given the AI prediction suggests that cases should either be decided by the AI alone or by the radiologist alone. Only a small share of cases are optimally delegated to radiologists with access to AI. In other words, we find that radiologists should work *next to* as opposed to *with* AI. To the extent that expert decision makers generally under-respond to information other than their own (Conlon et al., 2022) and incorporating additional information is cognitively costly, these insights may hold in other settings where experts' main job is a classification task.

There are several important considerations that are outside the scope of this work. One question motivated by the unrealized potential gains of AI assistance we find is whether the benefits from AI-specific training for radiologists and/or experience with AI are large. Answering these questions requires different experimental designs or longer-run studies. Another important consideration is that the organization of human-AI collaboration may influence experts' incentives, to which they may respond strategically. Finally, any interaction with AI will be mediated by organizational and regulatory incentives. Organizations may set guidelines on how to use AI or provide feedback, and regulations may influence liability implications. These issues are interesting avenues for future work.

AI continues to evolve rapidly. This development is increasingly driven by large corporate labs (Heston and Zwetsloot, 2020), which have resources that are unaffordable to academic institutions. For these and other reasons, economists are unlikely to have a major role in the technical development of AI tools. Our comparative advantage lies in studying how humans interact with these tools and thereby help shape the institutions that guide their use to ensure that this development is beneficial to society. Empirical analysis is a particularly useful tool in this endeavor, especially if the algorithms themselves are a black-box and cannot be understood from first principles.

References

Jason Abaluck, Leila Agha, Chris Kabrhel, Ali Raja, and Arjun Venkatesh. The determinants of productivity in medical testing: Intensity and allocation of care. *Am. Econ. Rev.*, 106 (12):3730–3764, December 2016.

Daron Acemoglu and Simon Johnson. Power and progress: Our Thousand-Year struggle over technology and prosperity. *Public Affairs, New York*, 2023.

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

Ajay Agrawal, Joshua S Gans, and Avi Goldfarb. Artificial intelligence: The ambiguous labor market impact of automating prediction. *J. Econ. Perspect.*, 33(2):31–50, May 2019.

Jong Seok Ahn, Shadi Ebrahimian, Shaunagh McDermott, Sanghyup Lee, Laura Naccarato, John F Di Capua, Markus Y Wu, Eric W Zhang, Victorine Muse, Benjamin Miller, Farid Sabzalipour, Bernardo C Bizzo, Keith J Dreyer, Parisa Kaviani, Subba R Digumarthy, and Mannudeep K Kalra. Association of artificial Intelligence–Aided chest radiograph interpretation with reader performance and efficiency. *JAMA Netw Open*, 5(8):e2229289–e2229289, August 2022.

Eugenio Alberdi, Lorenzo Strigini, Andrey A Povyakalo, and Peter Ayton. Why are people's decisions sometimes worse with computer support? In *Lecture Notes in Computer Science*, Lecture notes in computer science, pages 18–31. Springer Berlin Heidelberg, Berlin, Heidelberg, 2009.

Donald W K Andrews and Biao Lu. Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econom.*, 101(1): 123–164, March 2001.

Victoria Angelova, Will Dobbie, and Crystal Yang. Algorithmic recommendations and human discretion, 2022.

Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. Is the most accurate AI the best teammate? optimizing AI for teamwork. *AAAI*, 35(13):11405–11414, May 2021.

Nicholas Barberis, Andrei Shleifer, and Robert Vishny. A model of investor sentiment. *J. financ. econ.*, 49(3):307–343, September 1998.

Dan Benjamin, Aaron Bodoh-Creed, and Matthew Rabin. Base-Rate neglect: Foundations and implications, 2019.

Daniel J Benjamin. Chapter 2 - errors in probabilistic reasoning and judgment biases. In *Handbook of Behavioral Economics: Applications and Foundations 1*, volume 2, pages 69–186. January 2019.

Dirk Bergemann and Stephen Morris. Information design: A unified perspective. *J. Econ. Lit.*, 57(1):44–95, March 2019.

David Blackwell. Equivalent comparisons of experiments. *Ann. Math. Stat.*, 24(2):265–272, 1953.

Erik Brynjolfsson and Tom Mitchell. What can machine learning do? workforce implications. 358(6370):1530–, 2017.

Erik Brynjolfsson, Daniel Rock, and Chad Syverson. Artificial intelligence and the modern productivity paradox: A clash of expectations and statistics. November 2017.

David C Chan, Matthew Gentzkow, and Chuan Yu. Selection with variation in diagnostic skill: Evidence from radiologists. *Q. J. Econ.*, 137(2):729–783, May 2022.

Amitabh Chandra and Douglas O Staiger. Identifying sources of inefficiency in healthcare. *Q. J. Econ.*, 135(2):785–843, May 2020.

Gary Charness, Uri Gneezy, and Vlastimil Rasocha. Experimental methods: Eliciting beliefs. *J. Econ. Behav. Organ.*, 189:234–256, September 2021.

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

Emily F Conant, Alicia Y Toledano, Senthil Periaswamy, Sergei V Fotin, Jonathan Go, Justin E Boatsman, and Jeffrey W Hoffmeister. Improving accuracy and efficiency with concurrent use of artificial intelligence for digital breast tomosynthesis. *Radiol Artif Intell*, 1(4):e180096, July 2019.

John J Conlon, Malavika Mani, Gautam Rao, Matthew W Ridley, and Frank Schilbach. Not learning from others. August 2022.

Janet Currie and W Bentley MacLeod. Diagnosing expertise: Human capital, decision making, and performance among physicians. *J. Labor Econ.*, 35(1), 2017.

David Danz, Lise Vesterlund, and Alistair J Wilson. Belief elicitation: Limiting truth telling with information on incentives. June 2020.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.*, 144(1):114–126, February 2015.

Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. *Manage. Sci.*, 64(3):1155–1170, March 2018.

Enke and Zimmermann. Correlation neglect in belief formation. *Rev. Econ. Stud.*, 2019.

Ed Felten, Manav Raj, and Robert Seamans. How will language modelers like ChatGPT affect occupations and industries? March 2023.

Edward W Felten, Manav Raj, and Robert Seamans. The occupational impact of artificial intelligence: Labor, skills, and polarization. September 2019.

Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Paul Fisher, Diane Wilson, Michael Fitzke, Mark Parkinson, Eric Horvitz, Kori Inkpen, and Besmira Nushi. Who goes first? influences of Human-AI workflow on decision making in clinical imaging. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '22, pages 1362–1374. Association for Computing Machinery, June 2022.

Martin Ford. *Rise of the Robots: Technology and the Threat of a Jobless Future*. Basic Books, May 2015.

Morgan R Frank, David Autor, James E Bessen, Erik Brynjolfsson, Manuel Cebrian, David J Deming, Maryann Feldman, Matthew Groh, José Lobo, Esteban Moro, Dashun Wang, Hyejin Youn, and Iyad Rahwan. Toward understanding the impact of artificial intelligence on labor. *Proc. Natl. Acad. Sci. U. S. A.*, 116(14):6531–6539, April 2019.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo Koch, Matthias Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe Kitamura, Marzyeh Ghassemi, and Errol Colak. Who should do as AI say? only non-task expert physicians benefit from correct explainable AI advice. June 2022.

Susanne Gaube, Harini Suresh, Martina Raue, Eva Lermer, Timo K Koch, Matthias F C Hudecek, Alun D Ackery, Samir C Grover, Joseph F Coughlin, Dieter Frey, Felipe C Kitamura, Marzyeh Ghassemi, and Errol Colak. Non-task expert physicians benefit from correct explainable AI advice when reviewing x-rays. *Sci. Rep.*, 13(1):1383, January 2023.

Avi Goldfarb, Bledi Taska, and Florenta Teodoridis. Could machine learning be a general purpose technology? a comparison of emerging technologies using data from online job postings. *Res. Policy*, 52(1):104653, January 2023.

David M Grether. Bayes rule as a descriptive model: The representativeness heuristic. *Q. J. Econ.*, 95(3):537–557, November 1980.

David M Grether. Testing bayes rule and the representativeness heuristic: Some experimental evidence. *J. Econ. Behav. Organ.*, 17(1):31–57, January 1992.

Dale Griffin and Amos Tversky. The weighing of evidence and the determinants of confidence. *Cogn. Psychol.*, 24(3):411–435, July 1992.

Grimon, Marie-Pascale, and Christopher Mills. The impact of algorithmic tools on child protection: Evidence from a randomized controlled trial. 2022.

Jonathan Gruber, Benjamin R Handel, Samuel H Kina, and Jonathan T Kolstad. Managing intelligence: Skilled experts and decision support in markets for complex products. 2021.

H Benjamin Harvey and Vrushab Gowda. How the FDA regulates AI. *Acad. Radiol.*, 27(1): 58–61, January 2020.

Roxanne Heston and Remco Zwetsloot. Mapping u.s. multinationals' global AI R&D activity, 2020.

Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo J W L Aerts. Artificial intelligence in radiology. *Nat. Rev. Cancer*, 18(8):500–510, August 2018.

Tanjim Hossain and Ryo Okui. The binarized scoring rule. *Rev. Econ. Stud.*, 80(3):984–1001, February 2013.

Nicholas C Hunt and Andrea M Scheetz. Using MTurk to distribute a survey or experiment: Methodological considerations. *Journal of Information Systems*, 33(1):43–65, March 2019.

Kosuke Imai, Zhichao Jiang, James Greiner, Ryan Halen, and Sooahn Shin. Experimental evaluation of Algorithm-Assisted human Decision-Making: Application to pretrial public safety assessment. December 2020.

Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, Jayne Seekins, David A Mong, Safwan S Halabi, Jesse K Sandberg, Ricky Jones, David B Larson, Curtis P Langlotz, Bhavik N Patel, Matthew P Lungren, and Andrew Y Ng. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 590–597, July 2019.

Alistair E W Johnson, Tom J Pollard, Lu Shen, Li-Wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. MIMIC-III, a freely accessible critical care database. *Sci Data*, 3:160035, May 2016.

Daniel Kahneman and Amos Tversky. On the psychology of prediction. *Psychol. Rev.*, 80 (4):237–251, July 1973.

Emir Kamenica and Matthew Gentzkow. Bayesian persuasion. *Am. Econ. Rev.*, 101(6): 2590–2615, October 2011.

Hyo Eun Kim, Hak Hee Kim, Boo Kyung Han, Ki Hwan Kim, Kyunghwa Han, Hyeonseob Nam, Eun Hye Lee, and Eun Kyung Kim. Changes in cancer detection and False-Positive recall in mammography using artificial intelligence: a retrospective, multireader study. *The Lancet Digital Health*, 2(3):e138–e148, March 2020.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ziad Obermeyer. Prediction policy problems. *Am. Econ. Rev.*, 105(5):491–495, May 2015.

Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. Human decisions and machine predictions. *Q. J. Econ.*, 133(1):237–293, August 2017.

Barnett S Kramer, Christine D Berg, Denise R Aberle, and Philip C Prorok. Lung cancer screening with low-dose helical CT: results from the national lung screening trial (NLST). *J. Med. Screen.*, 18(3):109–111, 2011.

Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. Towards a science of Human-AI decision making: A survey of empirical studies. December 2021.

Curtis P Langlotz. Will artificial intelligence replace radiologists? *Radiology: Artificial Intelligence*, 1(3):e190058, May 2019.

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

Robert Mccluskey, A Enshaei, and B A S Hasan. Finding the Ground-Truth from multiple labellers: Why parameters of the task matter. *ArXiv*, 2021.

Hussein Mozannar and David Sontag. Consistent estimators for learning to defer to an expert. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 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.

Justin G Norden and Nirav R Shah. What AI in health care can learn from the long road to autonomous vehicles. *NEJM Catalyst Innovations in Care Delivery*, 3(2), 2022.

Shakked Noy and Whitney Zhang. Experimental evidence on the productivity effects of generative artificial intelligence. March 2023.

Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future — big data, machine learning, and clinical medicine. *N. Engl. J. Med.*, 375(13):1216–1219, September 2016.

Serena Pacilè, January Lopez, Pauline Chone, Thomas Bertinotti, Jean Marie Grouin, and Pierre Fillard. Improving breast cancer detection accuracy of mammography with the concurrent use of an artificial intelligence tool. *Radiol Artif Intell*, 2(6):e190208, November 2020.

David M Panicek and Hedvig Hricak. How sure are you, doctor? a standardized lexicon to describe the radiologist's level of certainty. *AJR Am. J. Roentgenol.*, 207(1):2–3, July 2016.

Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, Jason Ni, Fidaa Wishah, Fred Wittber, David S Hong, Thomas J Wilson, Safwan Halabi, Sanjay Basu, Bhavik N

Patel, Matthew P Lungren, Andrew Y Ng, and Kristen W Yeom. Deep Learning-Assisted diagnosis of cerebral aneurysms using the HeadXNet model. *JAMA network open*, 2(6): e195600, June 2019.

Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, Curtis Langlotz, Edward Lo, Joseph Mammarappallil, A J Mariano, Geoffrey Riley, Jayne Seekins, Luyao Shen, Evan Zucker, and Matthew Lungren. Human–Machine partnership with artificial intelligence for chest radiograph diagnosis. *npj Digital Medicine*, 2(1):111, December 2019.

M Rabin. Inference by believers in the law of small numbers. *Q. J. Econ.*, 2002.

M Rabin and D Vayanos. The gambler's and hot-hand fallacies: Theory and applications. *Rev. Econ. Stud.*, 2010.

Matthew Rabin. Incorporating limited rationality into economics. *J. Econ. Lit.*, 51(2): 528–543, June 2013.

Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv*, March 2019.

Pranav Rajpurkar and Matthew P Lungren. The current and future state of AI interpretation of medical images. *N. Engl. J. Med.*, 388(21):1981–1990, May 2023.

Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P Lungren, and Andrew Y Ng. CheXNet: Radiologist-Level pneumonia detection on chest X-Rays with deep learning. (1711.05225), December 2017.

Pranav Rajpurkar, Jeremy Irvin, Robyn L Ball, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis P Langlotz, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, Francis G Blankenberg, Jayne Seekins, Timothy J Amrhein, David A Mong, Safwan S Halabi, Evan J Zucker, Andrew Y Ng, and Matthew P Lungren. Deep learning for chest radiograph diagnosis: A retrospective comparison of the CheXNeXt algorithm to practicing radiologists. *PLoS Med.*, 15(11):e1002686, November 2018.

Pranav Rajpurkar, Chloe O'Connell, Amit Schechter, Nishit Asnani, Jason Li, Amirhossein Kiani, Robyn L Ball, Marc Mendelson, Gary Maartens, Daniël J van Hoving, Rulan Griesel, Andrew Y Ng, Tom H Boyles, and Matthew P Lungren. CheXaid: Deep learning assistance for physician diagnosis of tuberculosis using chest X-Rays in patients with HIV. *npj Digital Medicine*, 3:115, December 2020.

Pranav Rajpurkar, Emma Chen, Oishi Banerjee, and Eric J Topol. AI in health and medicine. *Nat. Med.*, 28(1):31–38, January 2022.

Ashesh Rambachan. Identifying prediction mistakes in observational data, 2021.

Carlo Reverberi, Tommaso Rigon, Aldo Solari, Cesare Hassan, Paolo Cherubini, and Andrea Cherubini. Experimental evidence of effective human–AI collaboration in medical decision-making. *Sci. Rep.*, 12(1):1–10, September 2022.

Michael Allan Ribers and Hannes Ullrich. Machine predictions and human decisions with variation in payoff and skills: the case of antibiotic prescribing, 2022.

Andrew B Rosenkrantz, Tarek N Hanna, Scott D Steenburg, Mary Jo Tarrant, Robert S Pyatt, and Eric B Friedberg. The current state of teleradiology across the united states: A national survey of radiologists' habits, attitudes, and perceptions on teleradiology practice. *J. Am. Coll. Radiol.*, 16(12):1677–1687, December 2019.

Jarrel C Y Seah, Cyril H M Tang, Quinlan D Buchlak, Xavier G Holt, Jeffrey B Wardman, Anuar Aimoldin, Nazanin Esmaili, Hassan Ahmad, Hung Pham, John F Lambert, Ben Hachey, Stephen J F Hogg, Benjamin P Johnston, Christine Bennett, Luke Oakden-Rayner, Peter Brotchie, and Catherine M Jones. Effect of a comprehensive deep-learning model on the accuracy of chest x-ray interpretation by radiologists: a retrospective, multireader multicase study. *Lancet Digit Health*, 3(8):e496–e506, August 2021.

Victor S Sheng, Foster Provost, and Panagiotis G Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 614–622, New York, NY, USA, August 2008. Association for Computing Machinery.

Yongsik Sim, Myung Jin Chung, Elmar Kotter, Sehyo Yune, Myeongchan Kim, Synho Do, Kyunghwa Han, Hanmyoung Kim, Seungwook Yang, Dong-Jae Lee, and Byoung Wook Choi. Deep convolutional neural network–based software improves radiologist detection of malignant lung nodules on chest radiographs. *Radiology*, 294(1):199–209, January 2020.

Christopher A Sims. Implications of rational inattention. *J. Monet. Econ.*, 50(3):665–690, April 2003.

Akshay Smit, Saahil Jain, Pranav Rajpurkar, Anuj Pareek, Andrew Y Ng, and Matthew P Lungren. CheXbert: Combining automatic labelers and expert annotations for accurate radiology report labeling using BERT. April 2020.

Megan Stevenson and Jennifer L Doleac. Algorithmic risk assessment in the hands of humans. December 2019.

Yasasvi Tadavarthi, Brianna Vey, Elizabeth Krupinski, Adam Prater, Judy Gichoya, Nabile Safdar, and Hari Trivedi. The state of radiology AI: Considerations for purchase decisions and current market offerings. *Radiol Artif Intell*, 2(6):e200004, November 2020.

Matt Taddy. The technological elements of artificial intelligence. February 2018.

Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, John Paoli,

Susana Puig, Cliff Rosendahl, H Peter Soyer, Iris Zalaudek, and Harald Kittler. Human–computer collaboration for skin cancer recognition. *Nat. Med.*, 26(8):1229–1234, June 2020.

A Tversky and D Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, September 1974.

Thomas S Wallsten and Adele Diederich. Understanding pooled subjective probability estimates. *Math. Soc. Sci.*, 41(1):1–18, January 2001.

Michael Webb. The impact of artificial intelligence on the labor market. 158713(November), 2019.

S Kevin Zhou, Hayit Greenspan, Christos Davatzikos, James S Duncan, Bram Van Ginneken, Anant Madabhushi, Jerry L Prince, Daniel Rueckert, and Ronald M Summers. A review of deep learning in medical imaging: Imaging traits, technology trends, case studies with progress highlights, and future promises. *Proc. IEEE*, 109(5):820–838, May 2021.