Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging

RICCARDO FOGLIATO, Carnegie Mellon University, USA
SHREYA CHAPPIDI, University of Virginia, USA
MATTHEW LUNGREN, Stanford University, USA
MICHAEL FITZKE and MARK PARKINSON, Mars Digital Technologies, USA
DIANE WILSON and PAUL FISHER, Antech Imaging Services, USA
ERIC HORVITZ, KORI INKPEN, and BESMIRA NUSHI, Microsoft Research, USA

Details of the designs and mechanisms in support of human-AI collaboration must be considered in the real-world fielding of AI technologies. A critical aspect of interaction design for AI-assisted human decision making are policies about the display and sequencing of AI inferences within larger decision-making workflows. We have a poor understanding of the influences of making AI inferences available before versus after human review of a diagnostic task at hand. We explore the effects of providing AI assistance at the start of a diagnostic session in radiology versus after the radiologist has made a provisional decision. We conducted a user study where 19 veterinary radiologists identified radiographic findings present in patients' X-ray images, with the aid of an AI tool. We employed two workflow configurations to analyze (i) anchoring effects, (ii) human-AI team diagnostic performance and agreement, (iii) time spent and confidence in decision making, and (iv) perceived usefulness of the AI. We found that participants who are asked to register provisional responses in advance of reviewing AI inferences are less likely to agree with the AI regardless of whether the advice is accurate and, in instances of disagreement with the AI, are less likely to seek the second opinion of a colleague. These participants also reported that the AI advice to be less useful. Surprisingly, requiring provisional decisions on cases in advance of the display of AI inferences did not lengthen the time participants spent on the task. The study provides generalizable and actionable insights for the deployment of clinical AI tools in human-in-the-loop systems and introduces a methodology for studying alternative designs for human-AI collaboration. We make our experimental platform available as open source to facilitate future research on the influence of alternate designs on human-AI workflows.

CCS Concepts: • **Computing methodologies** → Machine Learning; • **Human-centered computing** → *Human computer interaction (HCI)*; Empirical studies in interaction design; • **Applied computing** → Life and medical sciences.

Additional Key Words and Phrases: human-AI collaboration, decision making, clinical imaging, anchoring bias

ACM Reference Format:
Riccardo Fogliato, Shreya Chappidi, Matthew Lungren, Michael Fitzke, Mark Parkinson, Diane Wilson, Paul Fisher, Eric Horvitz, Kori Inkpen, and Besmira Nushi. 2022. Who Goes First? Influences of Human-AI Workflow on Decision Making in Clinical Imaging. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3531146.3533193

1 INTRODUCTION

We explore the influences of the sequencing of the availability of AI inferences on human decision making in a clinical imaging setting. We assess whether eliciting initial diagnoses from the participant before revealing the AI recommendation influences their final decisions and overall usage of AI inferences. In the study, we

FAccT '22, June 21–24, 2022, Seoul, Republic of Korea

ACM ISBN 978-1-4503-9352-2/22/06.
https://doi.org/10.1145/3531146.3533193

experiment with two human-AI collaboration workflow configurations. In the *one-step workflow*, participants were asked to identify radiographic findings given AI inferences and X-ray images at the same time. In the *two-step workflow*, participants were presented with AI inferences only after they had made a provisional decision. The two workflows reflect distinct approaches to interleaving AI assistance with human decision making. In practice, the joint presentation of the radiographic findings and inferential analysis at the start provides a more comprehensive set of information sources to the decision maker. However, as AI inferences may be erroneous, there are concerns about the possibility of unwanted anchoring that could lower the team's performance [4, 10, 29, 73].

Beyond focusing on performance, deployments must also consider preferences of human decision makers about alternate workflows when it comes to usability and adoption. While there is enthusiasm about bringing AI inferences into practice, resistance has been noted to AI assistance, with basis in multiple factors, including changes in established patterns of practice and aversion to automation [11, 21, 33]. For example, in high-stakes domains where human decision makers are the ultimate decision makers, there may be concerns regarding the influence of AI assistance on predictive performance, effort, and productivity. To understand the multifaceted impact of these tools, we measured (i) anchoring effects, (ii) human-AI team diagnostic performance and agreement, (iii) time spent and confidence in decision making, and (iv) perceived usefulness of the AI inferences.

We examined the influence of alternate human-AI workflows on a clinical imaging task with 19 veterinarian radiologist participants based at Mars, a pet healthcare company. The radiologists were asked to inspect and identify 33 different findings in X-ray images from real-world cases that had come to the company. In both workflows, the AI assistance consisted of binary AI inferences (finding present versus absent) on each finding along with the respective confidence scores. AI inferences were obtained from an ensemble machine learning model that is under consideration by Mars for deployment. The study was conducted via a web-based experimentation platform (Figure 1) where participants could inspect X-ray images, register their findings, and review AI inferences.

Key findings from the study demonstrate that alignment between participants' diagnoses and AI inferences is strongest in the one-step workflow, where radiologists were presented with AI assistance at the beginning of the diagnostic sessions. The findings highlight a higher risk of anchoring in the hone-step AI workflow. While we had hypothesized an anchoring effect given the nature of the workflow, we found that anchoring effects were minimal for findings considered critical or life-threatening for the animal. Although the AI outperformed participants, anchoring effects only led to marginal gains in diagnostic performance due to over reliance on erroneous AI advice. From a productivity perspective, we found that the time spent in both workflows was comparable. We were surprised to found that participants in the two-step workflow rarely revised their provisional diagnoses when the AI inferences differed from their earlier assessment. In perceptions shared in a survey, participants in the one-step workflow expressed a sense that the AI increased their confidence and speed more than those in the two-step workflow, and rated AI inferences as more useful. We believe that our multi-dimensional analysis provides actionable insights on the fielding of AI assistance in clinical imaging domains, suggesting that automated inferences may be most beneficial to human decision making when it is least disruptive per being smoothly integrated into the flow of human cognitive processes.

In summary, we make the following contributions:

- We conduct a user study to investigate the influence of two human-AI workflows on the diagnoses made by expert veterinary radiologists with the aid of an AI diagnostic tool. The analysis investigates questions about the influence of the sequencing of AI inferences on key dimensions, including anchoring bias, diagnostic performance, agreement, time spent, and user satisfaction.
- Based on the study results, we derive and discuss a set of implementable takeaways for the deployment of AI tools in human-driven decision-making processes.
- We release as open source the experimental platform, which can be used for conducting human-in-the-loop user studies in clinical imaging. The code is available at http://aka.ms/Exp-HAIC.

The rest of the paper is organized as follows. In Section 2, we position and contrast our work with previous findings in the human-AI collaboration and decision-making literature. Section 3 contains details of the experimental setup and the platform. Section 4 describes the study findings. Sections 5 and 6 discuss takeaways and future research directions.

2 BACKGROUND

2.1 Analyses of Human-AI Teams

AI tools are being deployed to aid human decision makers in a variety of high-stakes domains including healthcare, criminal justice, child welfare, and hiring [18, 20, 70]. In medicine, the recent advent of deep learning methods has sparked enthusiasm about translating prototypes in practice [31, 37, 74, 82], with systems showing performance on par with experts on diagnostic tasks [8, 25, 35, 40, 60–62, 72, 83, 90]. The hope is that these tools will produce sizable gains in efficiency of human decision-making processes [5, 52, 55, 57, 64, 75, 88]. However, evidence to date suggests that their deployment of AI systems does not necessarily yield a uniform improvement over the status quo [6, 24, 78, 80].

The adoption of these tools has fostered a scholarly effort on designing human-AI collaborations for optimal team decision making [4, 41, 59, 89]. Past studies in this space analyze how user performance and trust are affected by the presence of AI explanations [4, 54, 69, 93], the perceived and communicated AI accuracy [54, 91], and model updates [3], among others. A common theme in this body of work is that, when building an AI intended to collaborate with a human, numerous details of human-centered design need to be considered, in contrast to the dominant focus of attention on AI accuracy. Our study contributes new insights about the importance of workflow configurations as part of designs for human-AI collaboration. In contrast to prior studies, we analyze the decision making of domain experts, rather than of laypersons, on a diagnostic task in the medical space.

Prior studies have also analyzed human-AI teams in the clinical imaging setting. The studies to date largely focus on a comparison of diagnostic performance of humans alone versus human-AI teams [7, 36, 48, 56, 79, 81, 87]. These analyses, for the most part, report that interactions with AI tools lead to gains in human diagnostic performance. A handful of studies have compared the influences of various types of AI assistance on decisions [12, 76, 84]. Results from these investigations indicate that AI tools appropriately designed to support decision makers can boost not only diagnostic accuracy but also self-reported confidence, while decreasing mental effort and the time spent on the task. At the same time, some of these analyses have also witnessed the pitfalls of AI adoption, including increases in the time spent on the task without corresponding gains in accuracy, reductions in diagnostic performance due to reliance on erroneous AI advice, and participants ignoring AI advice altogether [48, 56, 84]. These are notable limitations because the performance of machine-learned models often varies across types of instances and can degrade over time [19, 83]. Poor integration of AI in human decision-making workflows can hamper the adoption of the tool by real-world decision makers [49, 58]. Although our experiment cannot fully emulate the clinical setting in which radiologists operate, our analysis attempts to capture the impact of the workflows across various dimensions.

2.2 Workflow Considerations for Human-AI Teams

A critical aspect of how AI can influence decision making revolves around the bias of anchoring [85]. Multiple studies have demonstrated that people may give stronger weight to their assessments towards prior knowledge or analyses versus doing full revisions in light of new evidence [86]. We hypothesized that anchoring effects of the review of information would be stronger when presented early versus late in problem solving. Thus we expected that AI inferences presented at the same time as initial analysis would be more influential than when the inferences are presented after an initial assessment. Research on the explanation of AI inferences frames opportunities for further study of the influences of designs for workflow of human-AI collaboration, including

altering the timing of AI-assistance and forcing users to spend more time on instances where AI inferences present higher uncertainty [4, 10, 32, 67, 73].

Several studies on human-AI collaboration have focused specifically on anchoring and workflow orderings similar to those studied in our experiment. Green and Chen [34] find that requiring participants to register provisional predictions before AI recommendations are revealed results in marginal gains in overall predictive performance. In a similar experimental setup, Fogliato et al. [29] do not detect anchoring effects or differences in performance across workflows. Buçinca et al. [10] report lower reliance on erroneous AI advice in the two-step workflow; we find similar results. In distinction to the these prior studies, rather than studies with lay participants,we study the behaviors and perceptions of domain experts on the tasks they perform in the course of their professional work. Assessments such as ours are critical requirements for real-world deployments because experts may have deeply ingrained processes for decision making and thus may interact with AI tools differently from crowdworkers employed in most studies of human-AI interaction. For example, experts may be reluctant about reviewing and leveraging AI advice [14, 33], a phenomenon referred to as "algorithm aversion" [21].

Beyond anchoring effects, we need to consider the cognitive cost of different sequencing of information fusion and decision making. Psychologists have shown that decision makers seek to minimize cognitive effort based on considerations of the perceived costs and benefits of the mental effort associated with different strategies for coming to a decision [53]. In this realm of research, studies have identified challenges with cognitive costs of aggregating new evidence [9, 68] and with considering sets of alternatives [39]. The cognitive effort required in a two-step versus one-step workflow to consider new information and to re-evaluate prior assessments has conceptual links to studies on the costs associated with task switching, interruption, and recovery [15, 43, 46]. Cognitive costs of re-examination when new information becomes available can be viewed as analogous to interruption and recovery on the initial task with new information [42, 44]. Thus, the re-opening of a completed analysis, as required in a two-step workflow, will tend to increase the cognitive effort required for a decision.

In findings related to cognitive effort, research on "cognitive forcing" has explored methods for pushing human decision makers to spend more time with deliberating about problems [10, 32, 67, 73]. Work in this area includes making AI assistance only available upon request or employing a "slow algorithm" that loads while the user waits to input their decision. While these cognitive forcing functions were found to increase performance measures and decrease AI reliance, they did so at the expense of additional time required for decision making [30]. Findings from other studies indicate that it is difficult for humans to revise or reverse their decisions due to psychological phenomena of sunk cost effects [2, 23], cognitive dissonance [26], and confirmation bias [51, 65]. Moreover, Kirkebøen et al. [50] find that decision reversals are associated with higher levels of post-outcome regret despite improved outcomes or predictive performance.

3 METHODS

We now describe task, experimental design, procedures, measures, data analysis, and experimental platform. The study received IRB approval by Mars. All X-ray images in the study were drawn from past patient examinations.

3.1 Experimental Task

Task. Study participants were shown 40 X-ray images from individual veterinary patients, which were all dogs for study consistency. The images were divided into two series of 20 images to reduce the load of a single session. For each image, radiologists were asked to diagnose which of 33 pre-specified radiographic findings could be identified. As shown in Figure 1, the diagnostic task for each of the findings required the following:

- Estimating the likelihood that the finding was present in the X-ray.
- Assessing whether to call the finding as present or absent in the X-ray.
- Flagging whether a second opinion from a colleague was needed.

The questions were modeled after assessments that radiologists make in their daily jobs. Response to the first request to assess the likelihood of findings were provided via a 0%–100% slider with bins of 10%. The remaining assessments were input using yes/no radio buttons. All answers were initialized to 0% and "no" respectively.

When making diagnoses, participants were assisted by an AI diagnostic tool that estimated the likelihood of each finding being present in the X-ray image. See Fitzke et al. [27] for a detailed description of the tool. The estimates are an average of predictions generated by eight separate convolutional neural networks, trained on a large proprietary dataset. To convert the estimates produced by the AI tool into binary predictions, we adopted a threshold of 0.6 which corresponds to the choice made while deploying the model in production to help radiologists label new data for AI training purposes. This threshold was derived by calibrating each model to maximize its probability predictions at 0.5 with regards to Youden's J-Statistic [92] and then further calibrating the resulting ensemble to 0.6 due to its accuracy gains [27]. Participants had access to both the likelihood and the binary prediction of present versus absent generated by the AI tool for each finding, which in the remainder of the paper we call *AI confidence* and *AI flags* respectively.

Dataset. The X-ray images were obtained from a benchmark dataset previously annotated by 10 to 13 expert radiologists independently. The radiologists labeled the 33 radiographic findings. We constructed ground truth labels based on these diagnoses by considering the finding as present when at least half of the radiologists had identified it in the image. For our study, we obtained 40 X-ray images from this dataset by oversampling those with the lowest agreement among radiologists to boost the difficulty of the task and thus maximize the power of our analysis on anchoring effects. Fleiss' kappa, a measure of inter-rater reliability, in the ground truth annotations was 0.44, which indicates weak agreement among radiologists [63]. In our final sample, only about 7% of all findings were present according to the ground truth majority vote annotations (91 out of 40×33=1320).

3.2 Experimental Design

Treatments. To test the impact of workflow configurations on decision making, we employed a between-subjects design by assigning each of the participants to one of two workflow configurations. In the *one-step workflow*, the X-ray image and the AI inferences (AI confidence and binary estimate) were shown at the same time. In the *two-step workflow*, participants were asked to make provisional diagnoses before the AI inferences were revealed. After seeing the automated inferences, they were allowed to revise their initial diagnoses. The studied workflows represent easily implementable human-AI team configurations that the company is considering for deployment. All participants reviewed the same two series of 20 images. Within each series, the images were reshuffled in random order for every participant to avoid ordering effects.

Procedure. We now describe how participants navigated through the web-based experimental platform. Participants were first shown a consent form and asked to provide an identifier they had been assigned in order to preserve anonymity. Next, they followed a series of instructions that included information about the content of the task and a description of the AI diagnostic tool. Importantly, we clarified in layman's terms that the AI confidence may not reflect frequentist probabilities. At the end of the instructions, participants completed a screening test with 10 questions designed to ensure that they understood the task structure and the information provided by the AI. Participants were prompted to revise their responses until they answered all questions correctly. After taking the screening test, participants reviewed each of the 40 images, one at a time, in two 20-image sessions, using the interface shown in Figure 1. The X-ray image was shown on the left of the web page and participants could zoom in and out, change brightness and contrast, or enlarge it to full-screen. These operations are typically available during X-ray evaluations. Diagnoses for each of the 33 radiographic findings could be made using the UI controls within the stacked frames in the middle of the web page, which also contained the corresponding AI confidence. To help participants easily navigate through the findings, a navigation bar appeared on the right

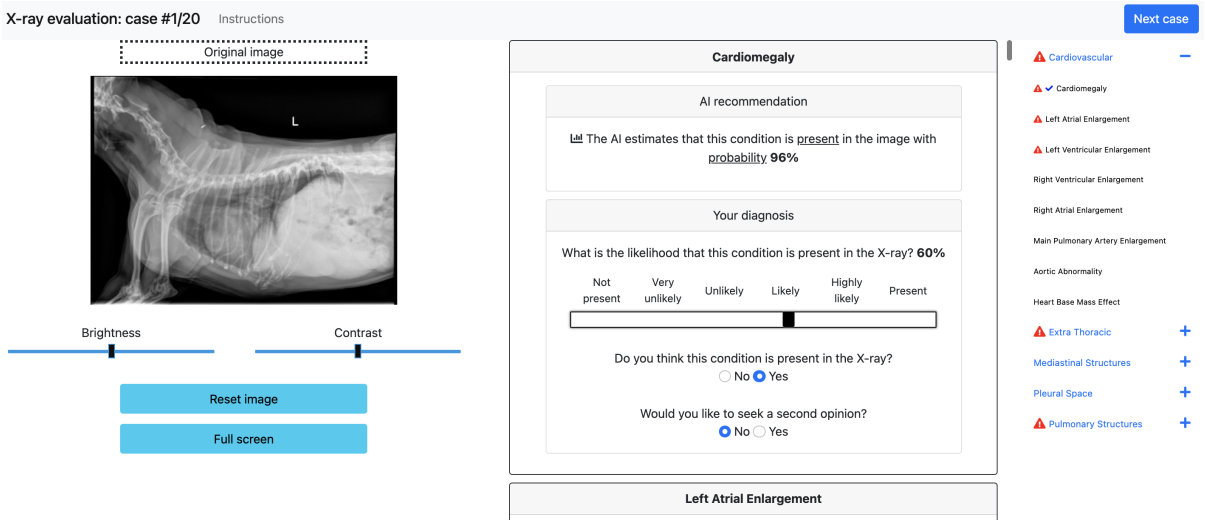

Fig. 1. Screenshot of the interface. On the left, an X-ray image of the thorax and abdomen of a dog is displayed. In the middle, the names of the radiographic findings with respective AI confidence are displayed in each of the boxes. Participants are asked to estimate the likelihood of the finding being present, whether to diagnose it as present or not, and whether they would seek a second opinion (default values are "0%", "No", and "No" respectively). Findings are grouped into macro-categories in the navigation bar on the right. When the AI flags a finding, a red triangle appears next to the name of both the finding and of the corresponding macro-category. If the participant identifies the finding as present, a check mark is shown.

of the page. Red triangles were shown next to the names of the findings flagged by the AI, i.e., those with AI confidence≥60%. Check mark symbols were displayed next to the findings identified by the participant in the image. AI confidence and flags were hidden from two-step workflow participants during their initial review of the image, and made visible only after they clicked a button in the top right corner of the interface. At the end of each image review, participants were asked whether the AI help had been useful, and then could proceed to the next image. Participants could not skip images or change diagnoses previously made. At the end of the experiment, they were asked to complete a questionnaire that we discuss in detail in Section 3.3.

Participants. A total of 24 veterinary radiologists employed at Mars were initially selected to participate in the study. Half of these radiologists were involved in data labeling of X-ray imagery as part of an ongoing organizational effort to embed machine learning in decision-making processes. The remaining radiologists had never interacted with AI tools nor had they done any labeling for AI training. Radiologists attended a one-hour orientation session during which the lead radiologist explained the purpose of the study (i.e., better understanding how to integrate AI in radiologists' decision-making processes), clarified the nature of the experimental task, and addressed questions and concerns. We assigned the radiologists to two experimental groups, one for each workflow configuration, balancing their years of experience and previous exposure to data labeling. In total, 19 participants started and completed the experimental task. The one- and two-step workflows had 11 and 8 submissions respectively, with five and three participants having prior exposure to data labeling for AI training. The median years of experience (9) was identical across workflows.

3.3 Measures and Statistical Analysis

Objective measures. We assessed the impact of workflow configurations through the following measures:

- Alignment between participants' diagnoses and AI inferences: Likelihood that the participants identify the same set of findings that are flagged by the AI.
- Diagnostic performance: Classification accuracy, false positive rate, false negative rate, and positive predicted values for both the AI and the participants' diagnoses.
- Inter-rater reliability: Fleiss' kappa for measuring agreement across participants [28].
- Time spent and confidence: Time spent reviewing each image, share of second opinions sought (a proxy for confidence), and likelihood estimates of the finding being present made by the participants.

In a separate survey we conducted, participants reported that reviews of complex cases encountered on the job take between 10 and 20 minutes. Thus, in the analysis of time on tasks we assumed that participants took breaks whenever they spent more than 15 minutes on a single image. Accordingly, we did not consider those observations (about 3% of all cases) in the analysis. The arbitrary choice of this threshold (instead of, say, 10 or 20 minutes) does not affect our study findings.

Taxonomy of findings. The lead radiologist determined whether each of the 33 findings satisfied the following five non-mutually exclusive criteria: (i) is critical, i.e., it requires immediate medical care and monitoring by healthcare professionals (vs. any other patient); (ii) is dangerous to overcall and treat if not actually present, i.e., it is important not to identify when absent; (iii) often requires a second opinion; (iv) has a vague definition, and (v) is common in animals and is often overlooked. We use these tags on findings later in Section 4 to perform a disaggregated analysis for critical vs. non-critical findings, to investigate diagnostic performance on findings that are dangerous to overcall, and to clarify workflow effects for findings that are expected to have a high disagreement (i.e., majority vote in ground truth being less reliable) versus those where lower disagreement is expected (i.e., majority vote in ground truth being more reliable). To this end, we would expect high disagreement among radiologists in findings that are at least in two of the categories (iii), (iv), and (v).

Subjective measures. We collected subjective measures of participants' confidence, diagnostic performance, and trust in the tool. After each review, participants were asked whether the AI inferences helped them to make their diagnoses. In addition, the final questionnaire elicited answers on seven-point Likert scales ranging from "strongly disagree" to "strongly agree" to assess the following measures:

- Workload: We inserted two questions related to the mental demand and frustration dimensions from the NASA-TLX study [38].
- Usefulness: We used two questions from the technology acceptance model (TAM) of Davis [16] related to gains in speed and diagnostic performance obtained by using the AI tool. Participants also reported on changes in confidence working alongside the AI. In addition, they could describe instances where the AI was most useful and least useful during their decision making via free-text responses.
- Future use: Participants indicated whether they would use a similar AI diagnostic tool in their daily jobs and could elaborate on their preference in an open-ended question.

To analyze differences in the ratings across workflows, we converted the Likert scale ratings into integers 1–7.

Statistical analysis. Prior to data collection, we planned to study the impact of workflow configurations on decision making by reporting summary statistics and the corresponding standard errors for each workflow, e.g., point estimate% [standard errors%]. These standard errors (e.g., of positive predicted values) are obtained via a nonparametric block bootstrap where the resampling is done at the participant's level, conditioning on their prior exposure to data labeling for AI training (see Section 3.8 of Davison and Hinkley [17]). We test the null hypothesis of independence of outcomes for the participant-level summary statistics and workflow configurations via rank-sum permutation tests, again conditioning on prior exposure to data labeling [45]. When relevant to the discussion, we report the corresponding one-sided (analysis of alignment between AI and participants) and two-sided (other analyses) p-values, considering a significance level of 0.1. We also examine participants' reliance

on AI flags by regressing their diagnoses on AI confidence and a dummy variable that indicates the presence of the AI flag, interacted with workflow configuration. This approach is inspired by regression discontinuity designs, a methodology popular in the econometric literature [1]. We fit the model via ordinary least squares and use sandwich standard errors clustered at the participant's level. Statistical significance of regression coefficients estimates is assessed via Wald tests. The plan for the analyses of the aforementioned measures at the aggregate level and conditional on AI inferences was laid out before running the experiment and motivated our data collection efforts. Prior to data analysis, we conducted a power analysis relative to our investigation of anchoring effects. We estimated power to be about 60% for a 2% standard deviation in the average agreement of participants with the AI and a difference of 1% in participants' agreement with the AI between workflows. The analysis based on the findings taxonomy represents a post-hoc investigation motivated by our study findings, which we decided to report because it reveals valuable insights into the nature of anchoring effects. We conducted an additional analysis using generalized linear mixed models that account for prior participant exposure to data labeling. The results obtained through this methodology are similar to those described in Section 4 and thus are omitted.

3.4 Experimental Platform

The platform was developed using the Python-based framework Django. The platform can accommodate future studies in similar domains by enabling researchers to bring in their own data sets of images, lists of ground truth diagnoses, and AI flag thresholds on algorithmic confidence. Either of the workflow configurations can be used for such studies. The platform logs relevant data on a per image basis regarding human diagnostic decisions, time elapsed, and responses to subjective questions. The platform also allows individuals to implement comprehension checks and collect data via surveys after the diagnostic tasks are completed.

4 RESULTS

4.1 Alignment Between Participants and AI

Result 1: Alignment between participants' diagnoses and AI inferences was highest in the one-step workflow, suggesting the influence of anchoring. This effect originated mostly from findings considered as non-critical for animal healthcare.

The final diagnoses made by participants matched the AI flags on 91% [standard error=1%] and 89% [1%] of the findings in the one- and two-step workflows, respectively. The alignment observed in the one-step workflow is significantly higher than in the two-step workflow (p-value is 0.04). Note that, of all <image, finding> pairs, only 11% were flagged as present by the AI, while 7% of them were marked as present by majority vote in the dataset. The low prevalence is explained by the fact that animals usually present only a few (and luckily not most) of the 33 findings in our list. Thus, all results in this section and the differences between the two workflows need to be interpreted with this consideration in mind.

When an AI flag was present for findings, participants in the one-step workflow were significantly more likely to identify the finding as present in the X-ray than their counterparts in the two-step workflow (in 71% [3%] and 65% [3%] of the findings, respectively; alignment is higher in the one-step workflow with p-value 0.07). When the AI flag was absent, participants in the one- and two-step workflows identified the finding in 7% [1%] and 8% [1%] of the cases (alignment is not significantly higher with p-value 0.18).

To better understand the impact of AI and anchoring biases in each workflow, Figure 2 shows the likelihood that participants identify a finding in the X-ray as a function of AI confidence. We are interested in the comparison of the discontinuity in the likelihood at AI confidence=0.6, the value used to determine whether the AI flag is shown. We observe that the estimated discontinuity for the participants' diagnoses in the one-step workflow is larger than for those in the two-step workflow (estimates of the dummy variable $\mathbb{1}$(AI confidence$\geq$0.6) are 0.24 [0.03] and 0.11 [0.03] in one- and two-step workflows, respectively; difference is statistically significant with

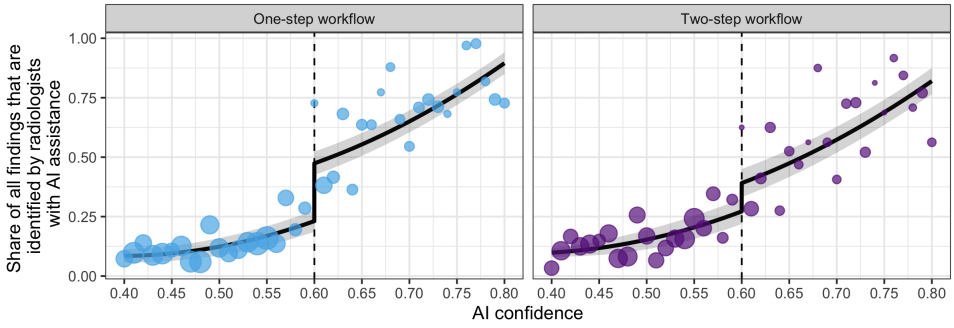

Fig. 2. Estimated probability that a participant would identify the radiographic finding in the X-ray as a function of AI confidence in the finding being present, for each workflow configuration. Each dot represents the share of findings that were identified for all findings with a certain AI confidence, across the entire set of images. The size of the dots is proportional to the number of findings. The parametric fits and corresponding 90% confidence intervals are represented by the solid black lines and gray shaded regions respectively. The magnitude of the estimated discontinuity at 0.6 in the one-step workflow is substantially larger than in the two-step workflow. This finding indicates that the presence of AI flags, which appeared for findings where AI confidence$\geq$ 0.6 (vertical dashed line), had a stronger influence on diagnoses made by one-step workflow participants. Further analysis reveals that this impact originates mostly from non-critical findings.

p-value<0.01). This means that the presence of an AI flag substantially increased the likelihood that participants would identify the finding compared to the one-step workflow. A nonparametric analysis further corroborates this result: Findings with AI confidence between 0.6 and 0.65 were identified by participants 48% [3%] and 40% [4%] of the times in one- and two-step workflows, respectively. Instead, those with AI confidence between 0.55 and 0.59 were identified 20% [3%] and 25% [4%] of the times respectively. These results indicate the presence of anchoring effects on AI flags in the one-step workflow.

A natural follow-up question is whether participants in the one-step workflow relied more on the AI uniformly across all of the findings. Using the taxonomy described in Section 3.3, we can investigate the phenomenon across two dimensions: the criticality of the finding and the difficulty of its interpretation. With respect to the criticality of findings (i.e., Question (i) in Section 3.3), we observe that anchoring effects were stronger on findings categorized as non critical. For example, when a non-critical finding was flagged by the AI, participants identified it in the X-ray in 76% [3%] and 65% [4%] of the cases in one- and two-step workflows respectively. When a critical finding was flagged instead, the respective shares were 67% [3%] and 65% [3%]. The discontinuity analysis delivers similar results. One explanation of this result is that participants might have put more effort into making these diagnoses, while pondering less the possible presence of AI flags and potentially more the value of AI confidence. We also unsurprisingly observe that anchoring effects are salient on findings where disagreement among radiologists is expected to be largest (based on Questions (iii, iv, v) in Section 3.3). When one of these finding was flagged by the AI, participants identified it in 73% [4%] and 64% [4%] of the cases in one- and two-step workflows respectively. The gap is virtually zero for the remaining not as difficult findings.

Finally, we briefly discuss whether and how participants in the two-step workflow revised their provisional diagnoses after observing AI inferences. In total, these participants changed their diagnoses on only 70 of the 10560 findings evaluated. This corresponds to 5% of all findings for which their initial diagnoses did not match the AI flags. The majority of these revisions (47) occurred for findings that were flagged by the AI but that the participants had not initially identified. Most of the remaining revisions (18) also happened in cases of disagreement between AI and provisional diagnoses, where the participant had initially identified the finding in

Table 1. Diagnostic performance of AI alone and of participants' diagnoses made with AI assistance [standard error %].

		Accuracy	False Positive Rate	False Negative Rate	Positive Predicted Values	% Predicted Positives
All findings	AI	94%	6%	18%	52%	11%
	One-step workflow	91% [1%]	8% [1%]	16% [3%]	42% [3%]	14% [1%]
	Two-step workflow	90% [1%]	9% [1%]	17% [3%]	39% [3%]	14% [1%]
Critical findings	AI	95%	5%	9%	55%	9%
	One-step workflow	94% [1%]	6% [1%]	14% [3%]	46% [5%]	10% [1%]
	Two-step workflow	92% [1%]	7% [1%]	15% [3%]	41% [4%]	12% [1%]
Findings dangerous to overcall	AI	96%	4%	10%	51%	8%
	One-step workflow	92% [1%]	7% [2%]	15% [4%]	35% [4%]	11% [2%]
	Two-step workflow	91% [1%]	9% [1%]	15% [4%]	32% [3%]	12% [1%]
Findings with lowest expected disagreement	One-step workflow, AI correct	96% [1%]	4% [1%]	11% [2%]	56% [4%]	8% [1%]
	Two-step workflow, AI correct	94% [1%]	6% [1%]	8% [2%]	46% [5%]	10% [1%]
	One-step workflow, AI incorrect	62% [2%]	41% [3%]	27% [4%]	32% [1%]	47% [3%]
	Two-step workflow, AI incorrect	66% [4%]	36% [4%]	25% [7%]	35% [3%]	44% [4%]

the image but the AI flag was absent. On the cases of initial disagreement, we could not detect any association between the tendency to revise and the criticality or difficulty of the finding.

4.2 Diagnostic Performance

Result 2: AI system outperformed participants across most of the performance metrics considered. Participants in the one-step workflow anchored more on the AI flags regardless of the AI accuracy, resulting in marginal gains in diagnostic performance when compared to the two-step workflow.

A critical dimension related to the impact of workflow configurations on human decision making is diagnostic performance. We conduct four related analyses of performance that consider various characteristics of the findings (Table 1). We now describe the key findings from each of these investigations in turn.

We start by considering all diagnoses made with AI assistance. The performance metrics relative to AI alone, one-step workflow, and two-step workflow participants are reported in the first three rows of Table 1. We observe that the AI outperformed participants on most of the metrics. Nonetheless, the only notable—yet not statistically significant difference—in performance across workflows is in the positive predicted values (42% vs. 39% in one- and two-step workflows respectively). Classification accuracy and false positive rate of the one-step workflow participants are also closer to those of the AI system, but the gains are minimal and our experiment is underpowered to detect such small variations.

Our second analysis focuses on critical findings. Similarly to the previous investigation, we find that, while the AI outperformed both groups of participants, those in the one-step workflow achieved slightly better performance, across all metrics. We repeat the analysis on findings that may be dangerous to overcall, for which making as few false positive diagnoses is crucial. We observe that the AI achieved the lowest false positive rate (4%), followed by those of participants in the one-step and two-step workflows (7% and 9% respectively).

Evaluations of diagnostic performance can be inherently problematic: radiologists often disagree on whether a certain finding is actually present, even in the original dataset. We mentioned this phenomenon when describing the ground truth annotations in Section 3.1. Thus, for some of the findings, a certain degree of disagreement between the diagnoses made by our study participants and ground truth should be expected. Our fourth analysis of performance focuses solely on the findings where we expect disagreement among radiologists to be lowest and thus ground truth annotations to be most reliable (again according to Questions (iii, iv, v) in Section 3.3). We find that diagnoses made in the one-step workflow achieved higher accuracy and lower false positive rates than those made in the two-step workflow. This mirrors our previous findings. However, given that the AI outperformed

participants by a considerable margin, why don't we observe larger gains in performance? One explanation is that even the diagnoses made in the one-step workflow did not always match AI flags. Moreover, these participants tended to agree more with the AI flags even when they were inaccurate. The last four rows of Table 1 report participants' diagnostic performance conditional on the accuracy of AI flags for these findings. We observe that, when AI advice was correct (e.g., a finding that was present was flagged), participants in the one-step workflow achieved better performance than those in the two-step workflow. When AI advice was wrong (e.g., a finding that was absent was flagged), they achieved worse performance, across all metrics. We observe an analogous phenomenon for critical findings with low expected disagreement, despite the minimal anchoring effects. These results demonstrate that the stronger alignment between AI and participants observed in the one-step workflow was not always warranted: Showing the AI flags directly made participants more likely to identify the finding not only when it was actually present but also also when ground truth indicated that it was not. These results explain the fact that anchoring led to only marginal gain in participants' overall performance.

4.3 Inter-rater Reliability

Result 3: Inter-rater reliability was highest for diagnoses made in the one-step workflow.

We have mentioned at several points in the paper that the diagnoses in the ground truth annotations from individual radiologists often differed. In our experiment, we expected the presence of the AI to affect agreement among radiologists differently across workflows. More specifically, we hypothesized that, as consequence of anchoring, (i) on findings where AI flags were accurate, agreement would be highest in the one-step workflow, i.e., one-step workflow participants would be more likely to make the same diagnoses; and (ii) on findings where AI flags were most likely inaccurate, agreement would be lowest in the one-step workflow. For (ii), we consider the findings on which we expected low disagreement among radiologists as described in Section 3.3. Thus, the hypothesized effects would run in different directions. We find that overall inter-rater reliability in the one-step workflow is higher than in the two-step workflow, with the respective estimates of Fleiss' kappa being 0.55 and 0.49. For (i), the inter-rater reliability measured on diagnoses made in the one-step workflow is higher compared to those in the two-step workflow; the respective estimates of Fleiss' kappa are 0.54 and 0.49. We find that the gap in kappas across workflows is largest on findings that are considered as non-critical for animal healthcare, for which we also observe the largest anchoring effects. We do not find evidence in support of (ii): Fleiss' kappas on findings with expected low disagreement are 0.37 and 0.35 in one- and two-step workflows respectively.

4.4 Time Spent on Decision Making and Confidence

Result 4: Time spent on the task did not differ across workflow configurations. In cases of disagreement with the AI, one-step workflow participants sought more often second opinions than their counterparts. Evidence suggests that they might have weighed AI inferences more meaningfully.

We expected that requiring participants to make provisional diagnoses before AI inferences were revealed would substantially slow down their decision making. The distribution of the time participants spent reviewing the images and making the diagnoses for each workflow is shown in Figure 3 (left). We observe that participants in the two-step workflow did not spend more time on the task than those in the one-step workflow, neither in terms of the average nor of the median times (the respective medians are 152 [standard error=20] and 139 [38] seconds, while averages are 189 [22] and 191 [36] seconds). We have identified two factors that may explain this surprising result. First, there exists a large variability in the time spent by participants, which calls for a larger sample size. Indeed, the fastest participant made the diagnoses in an average time of slightly more than one minute, while reviews took over six minutes to the slowest participant. A second hypothesis is that one-step workflow participants might have considered AI inferences more carefully.

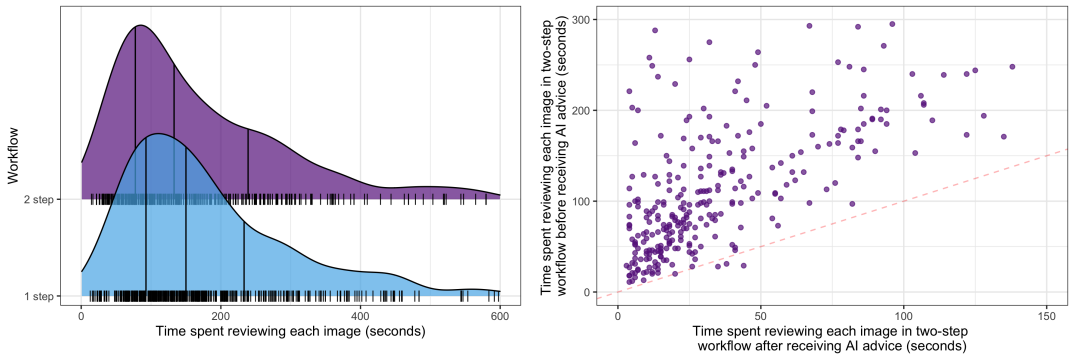

Fig. 3. Time spent by participants on decision making. The plot on the left shows the estimated densities of the time spent by participants reviewing the individual images and making the diagnoses, in each workflow. The vertical solid lines correspond to first, second (median), and third quartiles. The small vertical lines at the bottom represent individual observations, i.e., one image reviewed by one participant. The plot on the right shows the time participants in the two-step workflow spent before and after observing the AI inferences on each image (vertical and horizontal axes respectively). If the time spent did not differ across phases, the dots, which represent individual cases, would lie around the 45 degree dashed red line. For visualization purposes, we have limited the scales of the axes.

We can investigate the second hypothesis by examining participants' need of second opinions across workflows. As a reminder, our study participants were asked whether, were they to encounter the same patient in their daily job, they would seek the opinion of a colleague before making the final call on the diagnosis. This option was rarely chosen and the overall rates of second opinions were similar across workflows (about 1% of all findings evaluated). However, the two cohorts of participants tended to seek second opinions in different circumstances. On the one hand, one-step workflow participants sought second opinions *more* often in cases where they *disagreed* with the AI inferences. For findings that were identified by participants but were not flagged by the AI, the rates of second opinions were 11% [4%] and 6% [3%] for one- and two-step workflows respectively. For findings that were flagged by the AI but were not identified by participants, the respective rates were 2% [1%] and 1% [1%]. On the other hand, one-step workflow participants sought second opinions *less* often in cases where they *agreed* with AI inferences. This occurred in 1% [1%] and 3% [1%] of these findings that were flagged by both AI and participants in one- and two-step workflows respectively. The magnitude of these differences is substantially larger in case of findings that are critical or difficult to interpret according to our taxonomy (Question (iii) in Section 3.3). For example, for critical findings identified by the participant but not flagged by AI, second opinions were sought in 22% [7%] and 9% [6%] of the cases in one-and two-step workflows respectively. However, the rates of second opinions largely differed across participants as some of the participants never sought second opinions at all (those with more years of experience did so less often). Nonetheless, these empirical results appear to support the observation that participants in the one-step workflow considered the AI advice more meaningfully, and varied the need of second opinions according to their agreement with the recommendations. At the same time, we show in Section 4.1 that two-step workflow participants revised only a small number of their provisional diagnoses. Consistently, Figure 3 (right) highlights that these participants often spent a small amount of time reviewing the AI assistance (horizontal axis), supporting a similar interpretation as the analysis of second opinions.

It is possible that variations in confidence were reflected by the participants' subjective likelihood judgments of the findings being present. About three fourths of the collected estimates correspond to 0%, which was the default value. In all these cases, it is possible that participants believed that the finding was certainly absent or, in

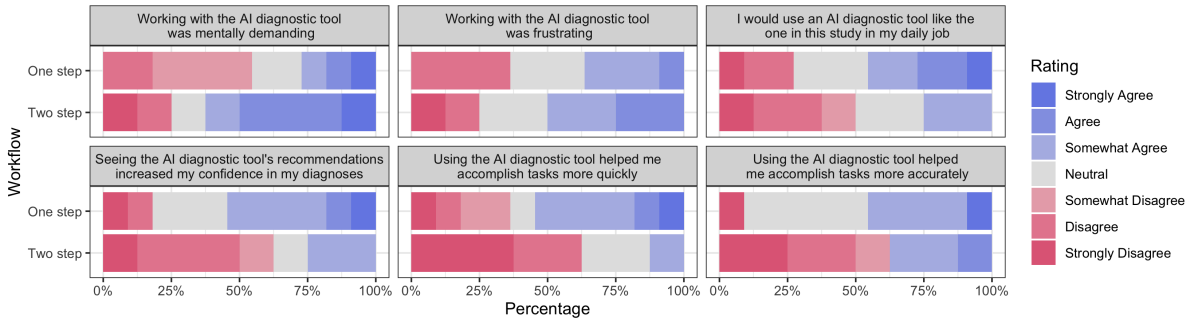

Fig. 4. Questions and corresponding answers in the final survey elicited on a seven-point Likert scale. The questions, reported in panel titles, concern perceived workload, future use of the AI tool, and perceived utility of the AI inferences.

the interest of time, that they didn't bother changing the default answer. Consequently, we focus our analysis only on findings that participants identified in the images. On these findings, the investigation of second opinions suggests that one-step workflow participants were less confident about their diagnoses compared to two-step workflow participants when the the AI flag was absent. However, we find that the average likelihood estimates for these findings are comparable across workflows (0.75 [0.02] and 0.77 [0.03] for one- and two-step workflows respectively, mean absolute differences between AI confidence and participants' estimates are 0.33 vs. 0.35). For findings that were instead flagged by the AI and were also identified by participants, two effects may be at play. On the one hand, one-step workflow participants may have anchored on the AI confidence. On the other hand, the analysis of second opinions suggests that their confidence may have been bolstered by the presence of the AI flag, yielding likelihood estimates higher than the AI's. The average values of the likelihood estimates are similar across workflows (0.89, [0.01] and 0.90 [0.02] in one- and two-step workflows, respectively). The discrepancy between these estimates and the AI confidence (0.15 and 0.16 respectively) is also comparable across workflows. Thus, the analysis suggests that the magnitude of subjective likelihood estimates did not vary across workflows.

4.5 Perceptions about AI inferences

Result 5: Participants in the one-step workflow rated the AI advice as more useful.

After reviewing each image and making their diagnoses, participants were asked whether the AI advice had been useful. Those assigned to the one-step workflow reported that the AI was useful for 36% [standard error=7%] of the images, whereas the two-step workflow participants found it useful in only 17% [4%] of the cases (p-value for hypothesis of independence is 0.11). The difference across workflows in the share of instances where the AI help was deemed useful is less than 10% for the first five images shown in the experiment and as large as 30% for the last five. Three questions in the final questionnaire can help us disentangle the AI's utility, or the lack thereof, into gains in accuracy, speed, and confidence in the diagnoses. Figure 4 shows the distribution of these ratings. We observe that participants assigned to the two-step workflow gave substantially lower ratings (or equivalently less utility) across all three dimensions. For gains in accuracy, the average ratings on the converted 1–7 Likert scale are 4.4 and 3.1 for one- and two-step workflows respectively. The respective ratings for confidence are 4.3 and 3.0 respectively. The difference relative to gains in speed is particularly striking: The average rating is 4.2 for one-step workflow participants and only 2.5 for those in the two-step workflow. This indicates that two-step workflow participants felt that the AI slowed down their decision making more than those in the one-step workflow. Two-sided p-values for tests of independence relative to confidence and speed are 0.1 and 0.05

respectively, while the p-value relative to accuracy is 0.26. In line with these results, we observe that two-step workflow participants reported the task to be slightly more stressful and the workload to be more demanding than their counterparts (average aggregate scores are 3.8 and 4.4 for one- and two-step workflows, with lower means indicating less stress and demand; however, we do not reject the null hypothesis of independence).

When asked for which diagnoses the AI advice had been *most* helpful, most of the participants (among those who answered) indicated that the AI helped them identify minor or incidental findings that would be less important in terms of decision making to the evaluating radiologist. This observation is consistent with the results of the quantitative analyses in Section 4.1 that highlighted more alignment between participants and AI inferences for non-critical findings. Participants also responded that the AI tool "was very helpful to reinforce [their] confidence", could be used when they were "on the fence about a finding instead of asking a colleague their opinion", and made them second guess their opinions by asking themselves whether they "could 'see' why it may have read it that way". When asked about the diagnoses for which the AI advice had been *least* helpful, participants indicated that for findings are erroneously flagged by the AI tool "it's an extra step to 'ignore' it" and that they spent "time searching for something that [they] ultimately decided isn't there". We did not identify notable differences in the answers across workflows.

Two-step workflow participants also appeared to be less willing to use this AI tool in their daily jobs (average ratings for one- and two-step workflows were 4.2 and 3.2 respectively), although we cannot reject the null hypothesis of independence. The participants who were reluctant to use the tool in the future expressed their frustration with the fact that the tool did not smoothly integrate into their workflows. Some reported that the AI tool was often (in their opinion) inaccurate and nudged them to spend extra time evaluating certain findings. This echoes the results described in the previous paragraph. One participant argued that, while the utility of the AI in its current form appeared to be limited, "if the tool were able to correctly interpret the images as [they] would and incorporate those findings into a report that [they] could then edit this could increase productivity".

5 DISCUSSION

Task realism. Several limitations need to be considered while interpreting the results of this study. First, the task setup did not include background clinical information on the patient (e.g., notes or historical background on the animal), which is typically available to the radiologist. Second, radiologists generally have access to multiple X-ray images and views from the patient. In our study, only one image was provided. The choice was motivated by the fact that the AI is only trained on individual images and does not leverage other clinical information. Having more views and clinical information available, radiologists may have exhibited different behaviors. Third, it is possible that some participants might pay more attention in their daily assessments than they did in our experiment given that they knew that these decisions would not impact animal treatment. Fourth, our study included only a short onboarding process covering the task and description of the AI tool. In real-world deployments, this phase should ideally be longer and provide more detailed information on the tool and its intended use [13, 47, 77].

Choices of interaction design. The platform entailed a series of design choices that require further analysis in future studies as they may have important effects in the interaction design of the workflow. Were the UI to change, for example through the removal of the AI flags from the navigation bar, our results could be affected. Examples of alternative designs include participants being asked to evaluate only a few of the findings at a time or review only the cases of disagreement with the AI inferences perhaps at a later stage. Further modifications of the interface could involve a dynamic selection of the workflow depending on findings based on the AI confidence. For instance, we observed that the agreement between participants and AI flags increased with AI confidence. Thus, it may be preferable for radiologists to first investigate findings where the uncertainty of AI inferences is highest. This proposal partially aligns with approaches explored by prior work, where human review was required

for the assessment of the most difficult tasks [59, 71]. Our findings also suggest that adjusting the threshold used to set AI flags by finding type (0.6 in our case) can influence final diagnostic decisions. Additionally, our participants used a novel experimental platform over two task sessions. Learning effects may have also impacted our experiment, particularly in terms of trust, ease of use, and time-related analyses. Such effects might have been more visible if the experiment had included more sessions.

Appropriate reliance. This study frames a question about the possibility of developing designs that could provide the best of both workflows: How might we obtain higher user satisfaction and engagement with the AI inferences seen in the one-step process while avoiding the increased tendency to anchor on erroneous AI inferences? We believe there is promise in studying modifications of the one-step workflow. For example, similarly to judges who deviate from sentencing guidelines per special considerations of the situation [66], radiologists could be asked to write down the reasons behind their diagnoses in case of disagreement with the AI inferences on a critical finding where AI confidence is far from the decision boundary. Alternatively, the opinion of a second radiologist may be required. This is, for example, what occurs in child welfare hotline screening decisions, where calls regarding children at high risk of out-of-home placement can only be screened out pending supervisor's approval [18]. At the same time, these potential modifications must be balanced with prior findings that cognitive forcing functions decrease trust and willingness to work with AI assistants [10]. Another possible strategy for trust calibration is represented by model explanations, beyond the likelihood estimates presented as AI confidence in our work. This direction has been explored by past work and holds potential in the clinical imaging context [22, 86].

6 CONCLUSIONS

We evaluated two methods for integrating AI inferences about radiographic findings into the workflows of veterinary radiologists, seeking to understand how the different approaches influence decision making. Our findings revealed that radiologists' diagnoses were more aligned with AI advice when it was shown immediately than in workflows where AI inferences were displayed after the radiologist had rendered a provisional assessment. The alignment, however, was similar across workflows for findings that were considered to be critical for the animal. Diagnoses made in the one-step workflow were characterized by marginal gains in diagnostic performance and higher levels of inter-rater reliability compared to those in the two-step workflow. Radiologists in the one-step workflow more frequently sought second opinions in cases of disagreement with the AI than in the two-step workflow and rated the AI tool was rated as more useful.

These results suggest that the one-step workflow can be meshed more smoothly with the current decision-making processes of radiologists than the two-step workflow. The dissatisfaction with AI assistance observed in the two-step workflow may be explained by the costs of task switching, interruption, and recovery described in Section 2. Adding a second step requires radiologists to stop, reassess, and reaffirm the diagnoses they had just completed. As a result, the AI advice shown after an analysis by the radiologists appears more prone to being disregarded. However, alignment between AI and radiologists was stronger in the one-step workflow even when AI inferences were inaccurate, suggesting increases in inappropriate reliance and anchoring on AI inferences.

At the highest level, our experiment demonstrates the importance of interaction design for clinical AI systems. We have explored a fundamental dimension of human-AI workflow, considering the effects of whether the AI inferences are made available immediately or following an AI-free analysis. Much remains to be explored.

ACKNOWLEDGMENTS

This study would not have been possible without the contribution and expertise of the veterinarian radiologists at Antech Imaging Services. The authors also thank Paul Koch for his help with the management of the infrastructure for the experiment, Lisa Ziemer for her support, and the anonymous reviewers for their valuable suggestions.

REFERENCES

[1] Joshua D Angrist and Jörn-Steffen Pischke. 2008. *Mostly harmless econometrics*. Princeton university press.

[2] Hal R Arkes and Catherine Blumer. 1985. The psychology of sunk cost. *Organizational Behavior and Human Decision Processes* 35, 1 (Feb 1985), 124–140. https://doi.org/10.1016/0749-5978(85)90049-4

[3] Gagan Bansal, Besmira Nushi, Ece Kamar, Daniel S Weld, Walter S Lasecki, and Eric Horvitz. 2019. Updates in human-AI teams: Understanding and addressing the performance/compatibility tradeoff. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2429–2437.

[4] Gagan Bansal, Tongshuang Wu, Joyce Zhou, Raymond Fok, Besmira Nushi, Ece Kamar, Marco Tulio Ribeiro, and Daniel Weld. 2021. Does the whole exceed its parts? the effect of ai explanations on complementary team performance. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[5] Mohsen Bayati, Mark Braverman, Michael Gillam, Karen M. Mark, George Ruiz, Mark S. Smith, and Eric Horvitz. 2014. Data-Driven Decisions for Reducing Readmissions for Heart Failure: General Methodology and Case Study. *PLOS One Medicine* 9, 10 (2014), 1–9.

[6] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M Vardoulakis. 2020. A human-centered evaluation of a deep learning system deployed in clinics for the detection of diabetic retinopathy. In *Proceedings of the 2020 CHI conference on human factors in computing systems*. 1–12.

[7] Nicholas Bien, Pranav Rajpurkar, Robyn L Ball, Jeremy Irvin, Allison Park, Erik Jones, Michael Bereket, Bhavik N Patel, Kristen W Yeom, Katie Shpanskaya, et al. 2018. Deep-learning-assisted diagnosis for knee magnetic resonance imaging: development and retrospective validation of MRNet. *PLoS medicine* 15, 11 (2018), e1002699.

[8] Titus J Brinker, Achim Hekler, Alexander H Enk, Joachim Klode, Axel Hauschild, Carola Berking, Bastian Schilling, Sebastian Haferkamp, Dirk Schadendorf, Stefan Fröhling, et al. 2019. A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task. *European Journal of Cancer* 111 (2019), 148–154.

[9] Arndt Bröder and Stefanie Schiffer. 2003. Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. *Journal of Experimental Psychology: General* 132, 2 (2003), 277.

[10] Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–21.

[11] Jason W Burton, Mari-Klara Stein, and Tina Blegind Jensen. 2020. A systematic review of algorithm aversion in augmented decision making. *Journal of Behavioral Decision Making* 33, 2 (2020), 220–239.

[12] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. 2019. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[13] Carrie J Cai, Samantha Winter, David Steiner, Lauren Wilcox, and Michael Terry. 2019. " Hello AI": Uncovering the Onboarding Needs of Medical Practitioners for Human-AI Collaborative Decision-Making. *Proceedings of the ACM on Human-computer Interaction* 3, CSCW (2019), 1–24.

[14] Lingwei Cheng and Alexandra Chouldechova. 2022. Heterogeneity in Algorithm-Assisted Decision-Making: A Case Study in Child Abuse Hotline Screening. *arXiv preprint arXiv:2204.05478* (2022).

[15] Mary Czerwinski, Edward Cutrell, and Eric Horvitz. 2000. Instant messaging and interruption: Influence of task type on performance. In *OZCHI 2000 conference proceedings*, Vol. 356. 361–367.

[16] Fred D Davis. 1989. Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS quarterly* (1989), 319–340.

[17] Anthony Christopher Davison and David Victor Hinkley. 1997. *Bootstrap methods and their application*. Number 1. Cambridge university press.

[18] Maria De-Arteaga, Riccardo Fogliato, and Alexandra Chouldechova. 2020. A case for humans-in-the-loop: Decisions in the presence of erroneous algorithmic scores. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[19] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. 2020. AI for radiographic COVID-19 detection selects shortcuts over signal. *medRxiv* (2020).

[20] Sarah L Desmarais, Samantha A Zottola, Sarah E Duhart Clarke, and Evan M Lowder. 2020. Predictive validity of pretrial risk assessments: A systematic review of the literature. *Criminal Justice and Behavior* (2020), 0093854820932959.

[21] Berkeley J Dietvorst, Joseph P Simmons, and Cade Massey. 2015. Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* 144, 1 (2015), 114.

[22] William K Diprose, Nicholas Buist, Ning Hua, Quentin Thurier, George Shand, and Reece Robinson. 2020. Physician understanding, explainability, and trust in a hypothetical machine learning risk calculator. *Journal of the American Medical Informatics Association* 27, 4 (2020), 592–600.

[23] Markus Domeier, Pierre Sachse, and Bernd Schäfer. 2018. Motivational Reasons for Biased Decisions: The Sunk-Cost Effect's Instrumental Rationality. *Frontiers in Psychology* 9 (May 2018), 815. https://doi.org/10.3389/fpsyg.2018.00815

[24] Pouyan Esmaeilzadeh, Murali Sambasivan, Naresh Kumar, and Hossein Nezakati. 2015. Adoption of clinical decision support systems in a developing country: Antecedents and outcomes of physician's threat to perceived professional autonomy. *International journal of medical informatics* 84, 8 (2015), 548–560.

[25] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. 2017. Dermatologist-level classification of skin cancer with deep neural networks. *nature* 542, 7639 (2017), 115–118.

[26] Leon Festinger. 1959. Review of A Theory of Cognitive Dissonance. *The American Journal of Psychology* 72, 1 (1959), 153–155. https://doi.org/10.2307/1420234

[27] Michael Fitzke, Conrad Stack, Andre Dourson, Rodrigo Santana, Diane Wilson, Lisa Ziemer, Arjun Soin, Matthew P Lungren, Paul Fisher, and Mark Parkinson. 2021. RapidRead: Global Deployment of State-of-the-art Radiology AI for a Large Veterinary Teleradiology Practice. *arXiv preprint arXiv:2111.08165* (2021).

[28] Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin* 76, 5 (1971), 378.

[29] Riccardo Fogliato, Alexandra Chouldechova, and Zachary Lipton. 2021. The Impact of Algorithmic Risk Assessments on Human Predictions and its Analysis via Crowdsourcing Studies. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW2 (2021), 1–24.

[30] Shane Frederick. 2005. Cognitive Reflection and Decision Making. *Journal of Economic Perspectives* 19, 4 (December 2005), 25–42. https://doi.org/10.1257/089533005775196732

[31] Joseph Futoma, Sanjay Hariharan, and Katherine Heller. 2017. Learning to detect sepsis with a multitask Gaussian process RNN classifier. In *International Conference on Machine Learning*. PMLR, 1174–1182.

[32] Krzysztof Z Gajos and Lena Mamykina. 2022. Do People Engage Cognitively with AI? Impact of AI Assistance on Incidental Learning. In *27th International Conference on Intelligent User Interfaces*. 794–806.

[33] Susanne Gaube, Harini Suresh, Martina Raue, Alexander Merritt, Seth J Berkowitz, Eva Lermer, Joseph F Coughlin, John V Guttag, Errol Colak, and Marzyeh Ghassemi. 2021. Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* 4, 1 (2021), 1–8.

[34] Ben Green and Yiling Chen. 2019. The principles and limits of algorithm-in-the-loop decision making. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–24.

[35] Holger A Haenssle, Christine Fink, Roland Schneiderbauer, Ferdinand Toberer, Timo Buhl, Andreas Blum, A Kalloo, A Ben Hadj Hassen, Luc Thomas, A Enk, et al. 2018. Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of oncology* 29, 8 (2018), 1836–1842.

[36] Seung Seog Han, Ilwoo Park, Sung Eun Chang, Woohyung Lim, Myoung Shin Kim, Gyeong Hun Park, Je Byeong Chae, Chang Hun Huh, and Jung-Im Na. 2020. Augmented intelligence dermatology: deep neural networks empower medical professionals in diagnosing skin cancer and predicting treatment options for 134 skin disorders. *Journal of Investigative Dermatology* 140, 9 (2020), 1753–1761.

[37] Intisar Rizwan I Haque and Jeremiah Neubert. 2020. Deep learning approaches to biomedical image segmentation. *Informatics in Medicine Unlocked* 18 (2020), 100297.

[38] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[39] John R Hauser and Birger Wernerfelt. 1990. An evaluation cost model of consideration sets. *Journal of consumer research* 16, 4 (1990), 393–408.

[40] Achim Hekler, Jochen S Utikal, Alexander H Enk, Wiebke Solass, Max Schmitt, Joachim Klode, Dirk Schadendorf, Wiebke Sondermann, Cindy Franklin, Felix Bestvater, et al. 2019. Deep learning outperformed 11 pathologists in the classification of histopathological melanoma images. *European Journal of Cancer* 118 (2019), 91–96.

[41] Sophie Hilgard, Nir Rosenfeld, Mahzarin R Banaji, Jack Cao, and David Parkes. 2021. Learning representations by humans, for humans. In *International Conference on Machine Learning*. PMLR, 4227–4238.

[42] Eric Horvitz and J. Apacible. 2003. Learning and Reasoning about Interruption,. In *International Conference on Multimodal Interaction*. 20–27.

[43] Eric Horvitz, Carl Kadie, Tim Paek, and David Hovel. 2003. Models of attention in computing and communication: From principles to applications. *Commun. ACM* 46, 3 (2003), 52–59.

[44] Eric Horvitz, Paul Koch, and Johnson Apacible. 2004. BusyBody: Creating and fielding personalized models of the cost of interruption. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*. 507–510.

[45] Torsten Hothorn, Kurt Hornik, Mark A Van De Wiel, and Achim Zeileis. 2006. A lego system for conditional inference. *The American Statistician* 60, 3 (2006), 257–263.

[46] Shamsi T Iqbal and Eric Horvitz. 2007. Disruption and recovery of computing tasks: Field study, analysis, and directions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 677–686.

[47] Maia Jacobs, Jeffrey He, Melanie F. Pradier, Barbara Lam, Andrew C Ahn, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. Designing AI for Trust and Collaboration in Time-Constrained Medical Decisions: A Sociotechnical Lens. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

[48] Ayush Jain, David Way, Vishakha Gupta, Yi Gao, Guilherme de Oliveira Marinho, Jay Hartford, Rory Sayres, Kimberly Kanada, Clara Eng, Kunal Nagpal, et al. 2021. Development and Assessment of an Artificial Intelligence–Based Tool for Skin Condition Diagnosis by Primary Care Physicians and Nurse Practitioners in Teledermatology Practices. *JAMA network open* 4, 4 (2021), e217249–e217249.

[49] Saif Khairat, David Marc, William Crosby, and Ali Al Sanousi. 2018. Reasons for physicians not adopting clinical decision support systems: critical analysis. *JMIR medical informatics* 6, 2 (2018), e24.

[50] Geir Kirkebøen, Erik Vasaasen, and Karl Halvor Teigen. 2013. Revisions and Regret: The Cost of Changing your Mind. *Journal of Behavioral Decision Making* 26, 1 (2013), 1–12. https://doi.org/10.1002/bdm.756

[51] Joshua Klayman. 1995. *Varieties of Confirmation Bias*. Psychology of Learning and Motivation, Vol. 32. Academic Press, 385–418. https://doi.org/10.1016/S0079-7421(08)60315-1

[52] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human decisions and machine predictions. *The quarterly journal of economics* 133, 1 (2018), 237–293.

[53] Wouter Kool, Joseph T. McGuire, Zev B. Rosen, and Matthew M. Botvinick. 2010. Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General* 139, 4 (2010), 665–682. https://doi.org/10.1037/a0020198

[54] Vivian Lai and Chenhao Tan. 2019. On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*. 29–38.

[55] Dae Hyun Lee, Meliha Yetisgen, Lucy Vanderwende, and Eric Horvitz. 2020. Predicting severe clinical events by learning about life-saving actions and outcomes using distant supervision. *Journal of Biomedical Informatics* 107 (2020), 103425.

[56] Constance D Lehman, Robert D Wellman, Diana SM Buist, Karla Kerlikowske, Anna NA Tosteson, Diana L Miglioretti, Breast Cancer Surveillance Consortium, et al. 2015. Diagnostic accuracy of digital screening mammography with and without computer-aided detection. *JAMA internal medicine* 175, 11 (2015), 1828–1837.

[57] Geert Litjens, Clara I Sánchez, Nadya Timofeeva, Meyke Hermsen, Iris Nagtegaal, Iringo Kovacs, Christina Hulsbergen-Van De Kaa, Peter Bult, Bram Van Ginneken, and Jeroen Van Der Laak. 2016. Deep learning as a tool for increased accuracy and efficiency of histopathological diagnosis. *Scientific reports* 6, 1 (2016), 1–11.

[58] Marjolein Lugtenberg, Jan-Willem Weenink, Trudy van der Weijden, Gert P Westert, and Rudolf B Kool. 2015. Implementation of multiple-domain covering computerized decision support systems in primary care: a focus group study on perceived barriers. *BMC medical informatics and decision making* 15, 1 (2015), 1–11.

[59] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict responsibly: Improving fairness and accuracy by learning to defer. *Advances in Neural Information Processing Systems* 31 (2018).

[60] Anna Majkowska, Sid Mittal, David F Steiner, Joshua J Reicher, Scott Mayer McKinney, Gavin E Duggan, Krish Eswaran, Po-Hsuan Cameron Chen, Yun Liu, Sreenivasa Raju Kalidindi, et al. 2020. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology* 294, 2 (2020), 421–431.

[61] Hisaki Makimoto, Moritz Höckmann, Tina Lin, David Glöckner, Shqipe Gerguri, Lukas Clasen, Jan Schmidt, Athena Assadi-Schmidt, Alexandru Bejinariu, Patrick Müller, et al. 2020. Performance of a convolutional neural network derived from an ECG database in recognizing myocardial infarction. *Scientific reports* 10, 1 (2020), 1–9.

[62] Roman C Maron, Michael Weichenthal, Jochen S Utikal, Achim Hekler, Carola Berking, Axel Hauschild, Alexander H Enk, Sebastian Haferkamp, Joachim Klode, Dirk Schadendorf, et al. 2019. Systematic outperformance of 112 dermatologists in multiclass skin cancer image classification by convolutional neural networks. *European Journal of Cancer* 119 (2019), 57–65.

[63] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[64] Scott Mayer McKinney, Marcin Sieniek, Varun Godbole, Jonathan Godwin, Natasha Antropova, Hutan Ashrafian, Trevor Back, Mary Chesus, Greg S Corrado, Ara Darzi, et al. 2020. International evaluation of an AI system for breast cancer screening. *Nature* 577, 7788 (2020), 89–94.

[65] Raymond S. Nickerson. 1998. Confirmation bias: A ubiquitous phenomenon in many guises. *Review of General Psychology* 2, 2 (1998), 175–220. https://doi.org/10.1037/1089-2680.2.2.175

[66] Pennsylvania Commission on Sentencing. 2014. Part VIII. Criminal Sentencing. Chapter 303. Sentencing guidelines. https://perma.cc/KR6G-94L7

[67] Joon Sung Park, Rick Barber, Alex Kirlik, and Karrie Karahalios. 2019. A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–15.

[68] John W Payne, James R Bettman, and Eric J Johnson. 1988. Adaptive strategy selection in decision making. *Journal of experimental psychology: Learning, Memory, and Cognition* 14, 3 (1988), 534.

[69] Forough Poursabzi-Sangdeh, Daniel G Goldstein, Jake M Hofman, Jennifer Wortman Wortman Vaughan, and Hanna Wallach. 2021. Manipulating and measuring model interpretability. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–52.

[70] Manish Raghavan, Solon Barocas, Jon Kleinberg, and Karen Levy. 2020. Mitigating bias in algorithmic hiring: Evaluating claims and practices. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 469–481.

[71] Maithra Raghu, Katy Blumer, Greg Corrado, Jon Kleinberg, Ziad Obermeyer, and Sendhil Mullainathan. 2019. The algorithmic automation problem: Prediction, triage, and human effort. *arXiv preprint arXiv:1903.12220* (2019).

[72] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. 2017. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225* (2017).

[73] Charvi Rastogi, Yunfeng Zhang, Dennis Wei, Kush R Varshney, Amit Dhurandhar, and Richard Tomsett. 2020. Deciding Fast and Slow: The Role of Cognitive Biases in AI-assisted Decision-making. *arXiv preprint arXiv:2010.07938* (2020).

[74] Dezső Ribli, Anna Horváth, Zsuzsa Unger, Péter Pollner, and István Csabai. 2018. Detecting and classifying lesions in mammograms with deep learning. *Scientific reports* 8, 1 (2018), 1–7.

[75] Paisan Ruamviboonsuk, Jonathan Krause, Peranut Chotcomwongse, Rory Sayres, Rajiv Raman, Kasumi Widner, Bilson JL Campana, Sonia Phene, Kornwipa Hemarat, Mongkol Tadarati, et al. 2019. Deep learning versus human graders for classifying diabetic retinopathy severity in a nationwide screening program. *NPJ digital medicine* 2, 1 (2019), 1–9.

[76] Rory Sayres, Ankur Taly, Ehsan Rahimy, Katy Blumer, David Coz, Naama Hammel, Jonathan Krause, Arunachalam Narayanaswamy, Zahra Rastegar, Derek Wu, et al. 2019. Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. *Ophthalmology* 126, 4 (2019), 552–564.

[77] Mark Sendak, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. " The human body is a black box" supporting clinical decision-making with deep learning. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 99–109.

[78] Jennifer Skeem, Nicholas Scurich, and John Monahan. 2020. Impact of risk assessment on judges' fairness in sentencing relatively poor defendants. *Law and human behavior* 44, 1 (2020), 51.

[79] David F Steiner, Robert MacDonald, Yun Liu, Peter Truszkowski, Jason D Hipp, Christopher Gammage, Florence Thng, Lily Peng, and Martin C Stumpe. 2018. Impact of deep learning assistance on the histopathologic review of lymph nodes for metastatic breast cancer. *The American journal of surgical pathology* 42, 12 (2018), 1636.

[80] Megan Stevenson. 2018. Assessing risk assessment in action. *Minn. L. Rev.* 103 (2018), 303.

[81] Stuart A Taylor, Susan C Charman, Philippe Lefere, Elizabeth G McFarland, Erik K Paulson, Judy Yee, Rizwan Aslam, John M Barlow, Arun Gupta, David H Kim, et al. 2008. CT colonography: investigation of the optimum reader paradigm by using computer-aided detection software. *Radiology* 246, 2 (2008), 463–471.

[82] Eric J Topol. 2019. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine* 25, 1 (2019), 44–56.

[83] Philipp Tschandl, Noel Codella, Bengü Nisa Akay, Giuseppe Argenziano, Ralph P Braun, Horacio Cabo, David Gutman, Allan Halpern, Brian Helba, Rainer Hofmann-Wellenhof, et al. 2019. Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: an open, web-based, international, diagnostic study. *The Lancet Oncology* 20, 7 (2019), 938–947.

[84] Philipp Tschandl, Christoph Rinner, Zoe Apalla, Giuseppe Argenziano, Noel Codella, Allan Halpern, Monika Janda, Aimilios Lallas, Caterina Longo, Josep Malvehy, et al. 2020. Human–computer collaboration for skin cancer recognition. *Nature Medicine* 26, 8 (2020), 1229–1234.

[85] Amos Tversky and Daniel Kahneman. 1974. Judgment under Uncertainty: Heuristics and Biases. *Science* 185, 4157 (1974), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

[86] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[87] Guangyu Wang, Xiaohong Liu, Jun Shen, Chengdi Wang, Zhihuan Li, Linsen Ye, Xingwang Wu, Ting Chen, Kai Wang, Xuan Zhang, et al. 2021. A deep-learning pipeline for the diagnosis and discrimination of viral, non-viral and COVID-19 pneumonia from chest X-ray images. *Nature Biomedical Engineering* (2021), 1–13.

[88] Jenna Wiens, John Guttag, and Eric Horvitz. 2016. Patient Risk Stratification with Time-Varying Parameters: A Multitask Learning Approach. *Journal of Machine Learning Research* 17, 79 (2016), 1–23.

[89] Bryan Wilder, Eric Horvitz, and Ece Kamar. 2021. Learning to Complement Humans. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI'20)*. Article 212, 8 pages.

[90] Nan Wu, Jason Phang, Jungkyu Park, Yiqiu Shen, Zhe Huang, Masha Zorin, Stanisław Jastrzebski, Thibault Févry, Joe Katsnelson, Eric Kim, et al. 2019. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE transactions on medical imaging* 39, 4 (2019), 1184–1194.

[91] Ming Yin, Jennifer Wortman Vaughan, and Hanna Wallach. 2019. Understanding the effect of accuracy on trust in machine learning models. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–12.

[92] William J Youden. 1950. Index for rating diagnostic tests. *Cancer* 3, 1 (1950), 32–35.

[93] Yunfeng Zhang, Q Vera Liao, and Rachel KE Bellamy. 2020. Effect of confidence and explanation on accuracy and trust calibration in ai-assisted decision making. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 295–305.