An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task

HELIODORO TEJEDA LEMUS, University of California, USA
AAKRITI KUMAR, University of California, USA
MARK STEYVERS, University of California, USA

Humans use AI assistance in a wide variety of high- and low-stakes decision-making tasks today. However, human reliance on the AI's assistance is often sub-optimal. In this paper, we present an empirical account of human-AI assisted decision-making in a noisy image classification task. We conduct an experiment where people are assisted by an AI in classifying noisy images into 16 categories. We analyze participants' reliance on the classifier's assistance and the accuracy of human-AI assistance as compared to the human or AI working independently. We demonstrate that participants do not show automation bias which is a widely reported bias displayed by humans when assisted by AI. In this specific instance of AI-assisted decision-making, people are able to correctly override the AI's decision when needed and achieve close to optimal performance. We suggest that the reason for this discrepancy from previous research findings is because a) people are experts at classifying everyday images and have a good understanding of their ability in performing the task, b) people engage in the metacognitive act of deliberation when asked to indicate confidence in their decision, and c) feedback was provided after each trial which enable people to build good mental models of the AI.

Additional Key Words and Phrases: AI Assistance, Decision-Making, Automation Bias, Human-Subject Experiment, Image Classification

ACM Reference Format:
Heliodoro Tejeda Lemus, Aakriti Kumar, and Mark Steyvers. 2022. An Empirical Investigation of Reliance on AI-Assistance in a Noisy-Image Classification Task. 1, 1 (April 2022), 8 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 INTRODUCTION

From making mundane shopping choices to thinking through high-stakes medical scenarios, there has been a sharp increase in the deployment of AI-assistants to help humans make decisions [1, 5, 8, 14]. In line with the old adage, "two minds are better than one", such collaborative human-AI decision-making are expected to increase the efficacy of decisions. However, recent work shows mixed results: while some studies report that decisions made jointly by the human and AI are more effective than either the human or the AI working independently [12, 13, 17, 19], other studies highlight humans' sub-optimal use of AI-advice and explanations [2, 18, 22]. Many empirical investigations of joint human-AI decision-making have indicated that humans are susceptible to biases and errors when working with AI assistance. People may over- or under-rely on the AI's advice leading to sub-optimal performance. Over- or under-reliance on the AI's assistance indicates miscalibrated 'trust' on the part of the human. Trust commensurate to the AI agent's capabilities is critical to effective joint decision-making.

Authors' addresses: Heliodoro Tejeda Lemus, , htejeda@uci.edu, University of California, Irvine, USA; Aakriti Kumar, aakritk@uci.edu, University of California, Irvine, USA; Mark Steyvers, mark.steyvers@uci.edu, University of California, Irvine, USA.

Manuscript submitted to ACM

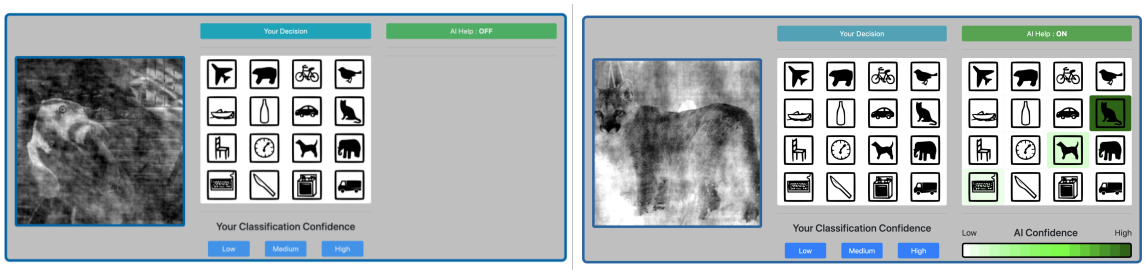
In this paper, we investigate people's reliance on an AI agent's assistance when the AI agent's advice is readily available to them when making a decision. We present data from a behavioral experiment where participants classify noisy images into one of 16 categories. The experiment has two within-subject conditions: one where participants can see an AI assistant's classification and confidence alongside the image, another where participants classify images without the AI classifier's help (control). We varied the accuracy of the AI assistants to look at differences in AI-assisted performance when the AI was better, similar, and worse than humans on the task. Figure 1 shows the experimental interface in both conditions.

Previous work has shown that a human's inclination to seek or incorporate advice is closely tied to their self-confidence in their decision [4, 9]. In our experiment, we capture participants' confidence ratings on each of their classifications. We use classification probabilities of the predicted class provided by the AI as a proxy for the its confidence in the classification [10]. We explore the reliance strategies of humans on the AI agent and accuracy of AI-assisted decision making. To preview the results, we observe that people's reliance on the AI assistant is close to optimal. We show that people do not over-rely on the AI even when their confidence in their own classifications is low. We posit that a useful characterisation of a human's reliance behavior on AI assistance needs to take into account not only the accuracy of the two agents, but also the confidence of the agents in their decisions. In our noisy-image classification task, we see that participants' performance improved when they were paired with any of three AI assistants with different levels of classification accuracy.

2 EXPERIMENTAL DESIGN

In the experiment, participants were tasked with classifying a total of 256 noisy images. On a subset of trials, AI assistance was provided (AI ON condition) while on another subset of trials, the AI assistance was turned off (AI OFF condition). Participants were instructed to classify all images as best they could and to leverage AI assistance (if provided) to optimize their performance. We use three classifiers of varying levels of accuracy to serve as assistants to the human participants. Each participant was assigned a single classifier level (A, B, or C) at the start of the experiment and would only be presented AI assistance from that particular classifier. One level of performance, classifier A, was set to perform below the baseline level of human performance. The second level of performance, classifier B, was set to be at roughly the baseline level. Finally, the third level of performance, classifier C, was created to be above the baseline level. All three models were fine-tuned in the same manner, training on all different phase noise levels all at once. However, to generate these different levels of performance, the models were fine-tuned for different periods of time.

Information about noise manipulation and classifier architecture as well as how they were trained can be found in [17]. The 256 trials were split in a block format in which AI assistance was turned on for 48 trials, followed by 16 trials without AI assistance. This process repeated 4 times to give a total of 192 tasks with AI assistance turned on while the remaining 64 were with AI assistance turned off. The figure below displays the experimental interface in both AI assistance conditions. The experiment had a three-column layout in which the leftmost column presented the noisy image that was to be classified. The middle column presented a grid of 16 category buttons for the participant to make their classification as well as three different submission buttons each representing a confidence level (low, medium, and high). Finally, the rightmost column was used for AI assistance. When AI assistance was turned off, this column displayed nothing. However, when AI assistance was turned on, there would be a grid of the 16 category options. Each of the 16 categories would be highlighted based on a gradient scale associated with the AI classifier prediction of that given category. The darker the hue of the highlighted category, the more confident the classifier was in that selection. For instances in which the classifier was extremely confident in a single category, there would only be one



(a) AI OFF condition

(b) AI ON condition

Fig. 1. Illustration of the behavioral experiment interface in both AI assistance conditions.

category highlighted with an extremely dark hue. Alternately, in instances where the classifier was not confident in a classification, there would be multiple categories highlighted with low hue levels. Participants were free to use the AI assistance to aid their final classification decision to the best of their abilities so as to optimize their own performance on the task. One important thing to note is that the same 256 images were presented to all participants in a random order. Having all participants classify the same 256 images allowed us to compare and contrast how participants classified a particular image with and without AI assistance. A total of 60 participants, 20 per classifier level, were recruited using Amazon Mechanical Turk.

We use the classification probabilities of the predicted class (highest probability class) as confidence values for the AI assistant. To simplify our analysis, we discretize these confidence values to match the human labels of low, medium, and high confidence. We set the AI confidence cutoffs of low, medium, and high to the intervals of (0.00-0.33], (0.33, 0.66], and (0.66, 1.00) respectively.

3 EMPIRICAL RESULTS

3.1 Is AI-assisted decision making more accurate than Human or the AI working independently?

Figure 2 shows the overall performance of the AI alone (classifier), humans alone (AI OFF), and the AI-assisted performance (AI ON) in the three conditions of the experiment. The second row of Figure 2 shows the change in accuracy in the AI ON condition versus the AI OFF condition. We see that humans are able to improve their performance across classifiers when aided by the AI. By looking at the second row of Figure 2, difference, we can more clearly see these improvements in accuracy across noise levels and how improvement grows with improving classifier accuracy. Figure 2 (A) is especially interesting because it indicates that humans are able to appropriately rely on the AI and improve their performance even when aided by an AI that has has worse accuracy than humans on average. This trend of improved accuracy of the human-AI assisted condition is consistent across the three classifiers. Participants in our experiment show appropriate reliance on the AI assistant and hence are able to improve their performance.

3.2 Are people's reliance strategies *optimal*?

We compare the observed strategy accuracy to the expected accuracy when an 'optimal' strategy is adopted. We define observed strategy as the strategy employed by participants in the AI ON condition. We define the optimal strategy as the best possible performance that may be achieved by combining the human and the AI's classifications. We simulate accuracy under the optimal strategy by marking the image as being classified correctly if either of the human or the AI

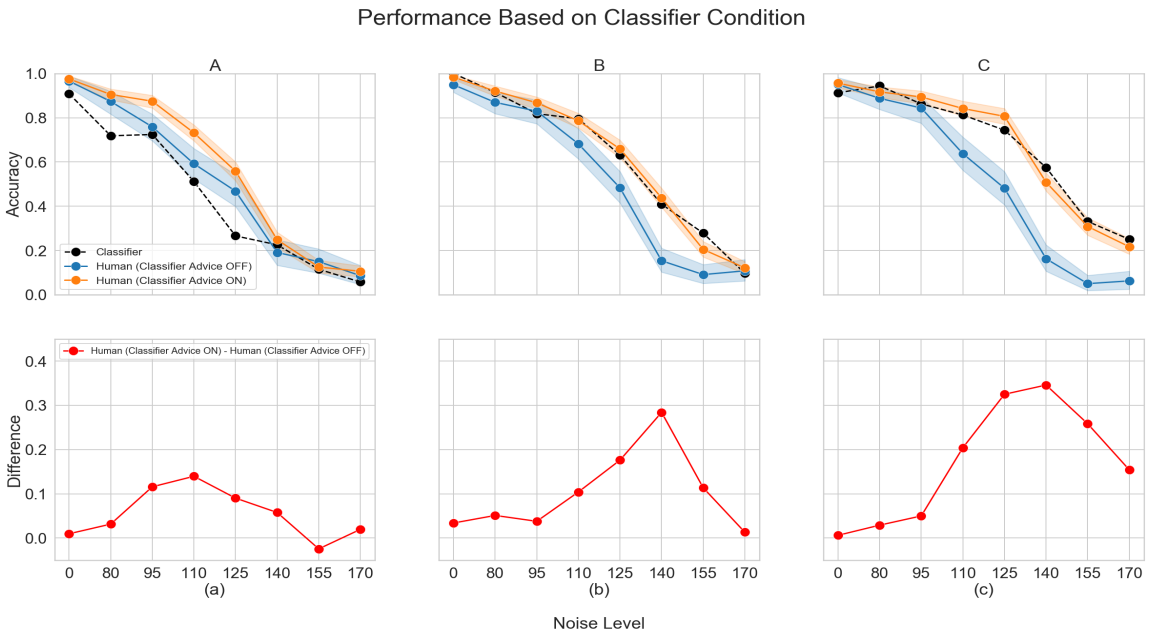


Fig. 2. Participant performance across noise levels. Row one shows performance on the task. Row two shows the difference in performance between AI ON - AI OFF conditions. Columns represents classifier levels (A, B, and C).

classify the image correctly. Figure 3 (a) shows the potential accuracy when adopting the optimal policy, (b) shows the observed accuracy in the AI ON condition, and (c) shows the difference in accuracy when following an optimal strategy as compared to the observed accuracy in our empirical data (Optimal - AI ON condition). We see that on average, people's accuracy is close to the expected accuracy under an optimal strategy. This trend is consistent across different levels of human and AI confidence.

3.3 Is people's behavior consistent with *automation bias*?

Automation bias is described as a human's tendency to over-rely on machine recommendations [7]. It has been widely reported as a bias displayed by humans [7, 11, 21]. In our experiment, the path of least resistance for a participant would be to agree with the AI's decision as it is always available in the AI ON condition trials. Hence, it is necessary to check for automation bias. We compare the AI ON condition's actual performance to a strategy that would describe automation bias in our system. Such a strategy would always select the AI's classification decision. Figure 4 (c) shows the expected difference in accuracy if participant's were to adopt a policy consistent with automation bias as compared to the policy observed in the AI ON condition of our empirical data. We see that people's policy is easily distinguishable from automation-bias. In cases where the human's confidence is low, we see that the accuracy as observed in the AI ON condition far exceeds the expected accuracy of the automation bias policy. This indicates that the images where the human had low confidence were also regions of low accuracy for the classifier. Humans still employed a strategy better than over-relying on the AI. Alternately, in instances where the human's confidence is high, we see that the accuracy as observed in the AI ON condition is comparable to the expected accuracy of the automation bias policy.

Fig. 3. Accuracy when adopting the (a) optimal policy, (b) the observed policy; and (c) the difference in accuracy when adopting the optimal policy versus the observed policy (Optimal - Observed) . Rows correspond to the three different classifiers A, B and C. Within each grid, we have human confidence (low, medium, high) on the x-axis and AI confidence (low, medium, high) on the y-axis.

4 DISCUSSION

This paper adds to a growing body of literature that investigates AI-assisted decision-making. Our empirical results reveal that in this image classification task, where people have a good understanding of their own ability and confidence on each trial, they are close to optimal in their adoption of the AI's advice. We show that performance in the AI ON condition does not deteriorate even in cases where the AI assistant has worse accuracy on the task than the human. We also find that people's reliance strategy on the AI in this task is not consistent with behavior that is associated with automation bias.

Most recent investigations of human decision-making with AI-assistance follow a judge-advisor system [15, 16] where humans are required to independently solve the task at hand before they are shown an AI assistant's recommendation [4, 9**?**]. Once shown the AI's advice, humans may update their final decision. Such experimental paradigms provide direct insights about a human's reliance behavior and make it easier for experimenters to disentangle the influence of the AI's advice on the final decision of the human. Also, deliberation and independent assessment of a problem have shown to help decrease over reliance on the AI [3]. However, this setup is somewhat artificial and incompatible with how AI assistants work in the real world. The ultimate aim of providing AI-assistance is to reduce the workload of humans and improve the accuracy of the joint decisions. We argue that our setup is a more natural way of incorporating

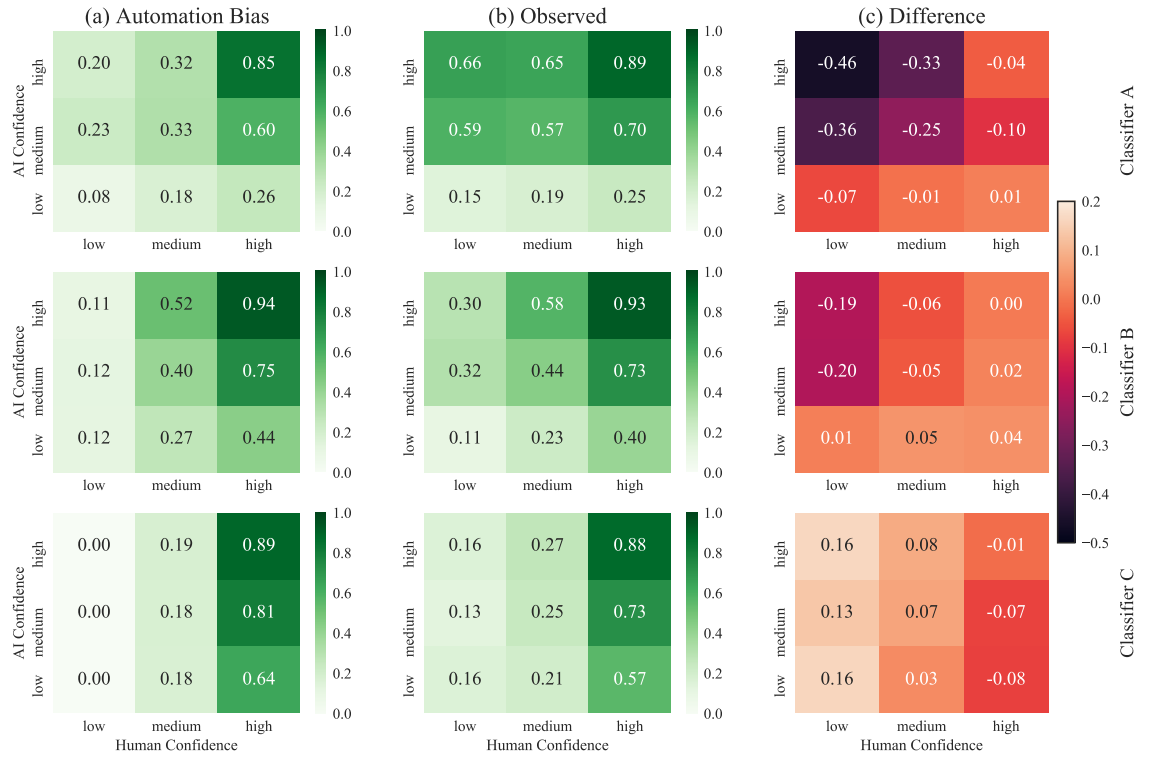


Fig. 4. Accuracy when adopting (a) a policy consistent with automation bias, (b) the observed policy (AI ON); and (c) the difference in accuracy when adopting the automation bias policy versus the observed policy in the AI ON condition (Automation Bias - Observed). Rows correspond to the three different classifiers A, B and C. Within each grid, we have human confidence (low, medium, high) on the x-axis and AI confidence (low, medium, high) on the y-axis.

AI assistance into everyday workflows. In our experiment, in the AI ON condition, the AI assistant's advice and its confidence in the advice is available to the human as soon as the task is presented. While this format of providing assistance may raise concerns about automation bias, we show that people adopted close to optimal strategies in our task. The AI OFF condition gives us information about the humans' independent classification judgement and confidence rating which we use to indirectly assess the influence of advice.

We believe that participants in our experiment were able to build close to optimal reliance strategies because of the following reasons. First, this is a simple task and most people are experts at identifying everyday objects. This enables people to have a good understanding of their own expertise and confidence on any presented image. Second, indicating confidence in a decision requires humans to employ a second-order metacognitive computation to evaluate their decision [6]. While this may not be an obvious 'cognitive forcing function' [3], prior work indicates that the mechanism humans employ to generate confidence ratings requires metacognitive deliberation about the strength of evidence available to make the decision [20]. Finally, in our experiment, people received feedback after each trial, which gave them the opportunity to learn about the AI assistant's accuracy and confidence calibration. We show that people were able to use feedback and build reasonable mental models of the AI assistant when paired with any of the three

classifiers of varying levels of accuracy. However, immediate feedback is not always possible in real-world scenarios. The impact of delayed feedback on the reliance behavior must be investigated in isolation.

In future work, we will present a cognitive model that can predict a human's classification decision and confidence rating on each trial based on information gathered about the human during the AI OFF condition. Such a model would allow us to infer the human's 'latent' decision to switch to the AI's recommendation without explicitly asking the human to provide an independent judgement. We believe that this work has important implications in designing better AI assistants and explanations. A key limitation of our work is that we used a low-stakes image classification task. Further work is needed to understand how performance varies with varying confidence of the agents and how this generalizes to other tasks.

5 AUTHORS AND AFFILIATIONS

REFERENCES

[1] Suresh Kumar Annappindi. 2014. System and method for predicting consumer credit risk using income risk based credit score. US Patent 8,799,150.

[2] 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.

[3] 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.

[4] Leah Chong, Guanglu Zhang, Kosa Goucher-Lambert, Kenneth Kotovsky, and Jonathan Cagan. 2022. Human confidence in artificial intelligence and in themselves: The evolution and impact of confidence on adoption of AI advice. *Computers in Human Behavior* 127 (2022), 107018. https://doi.org/10.1016/j.chb.2021.107018

[5] Steven E Dilsizian and Eliot L Siegel. 2014. Artificial intelligence in medicine and cardiac imaging: harnessing big data and advanced computing to provide personalized medical diagnosis and treatment. *Current cardiology reports* 16, 1 (2014), 1–8.

[6] Stephen M Fleming and Nathaniel D Daw. 2017. Self-evaluation of decision-making: A general Bayesian framework for metacognitive computation. *Psychological review* 124, 1 (2017), 91.

[7] Kate Goddard, Abdul Roudsari, and Jeremy C Wyatt. 2012. Automation bias: a systematic review of frequency, effect mediators, and mitigators. *Journal of the American Medical Informatics Association* 19, 1 (2012), 121–127.

[8] Nina Grgić-Hlača, Christoph Engel, and Krishna P Gummadi. 2019. Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–25.

[9] Aakriti Kumar, Trisha Patel, Aaron S Benjamin, and Mark Steyvers. 2021. Explaining Algorithm Aversion with Metacognitive Bandits. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 43.

[10] Vivian Lai, Chacha Chen, Q Vera Liao, Alison Smith-Renner, and Chenhao Tan. 2021. Towards a Science of Human-AI Decision Making: A Survey of Empirical Studies. *arXiv preprint arXiv:2112.11471* (2021).

[11] Raja Parasuraman, Robert Molloy, and Indramani L Singh. 1993. Performance consequences of automation-induced'complacency'. *The International Journal of Aviation Psychology* 3, 1 (1993), 1–23.

[12] Bhavik N Patel, Louis Rosenberg, Gregg Willcox, David Baltaxe, Mimi Lyons, Jeremy Irvin, Pranav Rajpurkar, Timothy Amrhein, Rajan Gupta, Safwan Halabi, et al. 2019. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. *NPJ Digital Medicine* 2, 1 (2019), 1–10.

[13] P Jonathon Phillips, Amy N Yates, Ying Hu, Carina A Hahn, Eilidh Noyes, Kelsey Jackson, Jacqueline G Cavazos, Géraldine Jeckeln, Rajeev Ranjan, Swami Sankaranarayanan, et al. 2018. Face recognition accuracy of forensic examiners, superrecognizers, and face recognition algorithms. *Proceedings of the National Academy of Sciences* 115, 24 (2018), 6171–6176.

[14] 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. 2020. CheXaid: deep learning assistance for physician diagnosis of tuberculosis using chest x-rays in patients with HIV. *npj Digital Medicine* 3, 1 (Sept. 2020). https://doi.org/10.1038/s41746-020-00322-2

[15] JA Sniezek and Timothy Buckley. 1989. Social influence in the advisor-judge relationship. In *Annual meeting of the Judgment and Decision Making Society, Atlanta, Georgia.*

[16] Janet A Sniezek, Gunnar E Schrah, and Reeshad S Dalal. 2004. Improving judgement with prepaid expert advice. *Journal of Behavioral Decision Making* 17, 3 (2004), 173–190.

[17] Mark Steyvers, Heliodoro Tejeda, Gavin Kerrigan, and Padhraic Smyth. 2022. Bayesian Modeling of Human-AI Complementarity in Image Classification. *Proceedings of the National Academy of Sciences* (2022).

[18] Sarah Tan, Julius Adebayo, Kori Inkpen, and Ece Kamar. 2018. Investigating human+ machine complementarity for recidivism predictions. *arXiv preprint arXiv:1808.09123* (2018).

[19] Darryl E Wright, Chris J Lintott, Stephen J Smartt, Ken W Smith, Lucy Fortson, Laura Trouille, Campbell R Allen, Melanie Beck, Mark C Bouslog, Amy Boyer, et al. 2017. A transient search using combined human and machine classifications. *Monthly Notices of the Royal Astronomical Society* 472, 2 (2017), 1315–1323.

[20] Nick Yeung and Christopher Summerfield. 2012. Metacognition in human decision-making: confidence and error monitoring. *Philosophical Transactions of the Royal Society B: Biological Sciences* 367, 1594 (2012), 1310–1321.

[21] Guanglu Zhang, Ayush Raina, Jonathan Cagan, and Christopher McComb. 2021. A cautionary tale about the impact of AI on human design teams. *Design Studies* 72 (2021), 100990.

[22] 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.