Effective Stereotypical Bias Detection: The Impact of Human-AI Collaboration Modes on Human Reliance on AI Recommendation

12-10-2024

Effective Stereotypical Bias Detection: The Impact of Human-AI Collaboration Modes on Human Reliance on AI Recommendation

Danlei Mou
University of Melbourne, dmou@student.unimelb.edu.au

Tingru Cui
University of Melbourne, tingru.cui@unimelb.edu.au

Katja Holtta-Otto
University of Melbourne, katja.holttaotto@unimelb.edu.au

Bo Du
Griffith University, bo.du@griffith.edu.au

Jiawei Tong
University of Melbourne, jiawei.tong.1@unimelb.edu.au

Effective Stereotypical Bias Detection: The Impact of Human-AI Collaboration Modes on Human Reliance on AI Recommendation

Research-in-progress

Danlei Mou
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
Email: dmou@student.unimelb.edu.au

Tingru Cui
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
Email: tingru.cui@unimelb.edu.au

Katja Holtta-Otto
Department of Mechanical Engineering
The University of Melbourne
Melbourne, Australia
Email: katja.holttaotto@unimelb.edu.au

Bo Du
Department of Business Strategy and Innovation
Griffith University
Brisbane, Australia
Email: bo.du@griffith.edu.au

Jiawei Tong
School of Computing and Information Systems
The University of Melbourne
Melbourne, Australia
Email: jiawei.tong.1@unimelb.edu.au

Abstract

As artificial intelligence (AI) increasingly supports human decision-making across various domains, from hiring processes to legal judgments, understanding how to optimize human-AI collaboration is crucial for developing trustworthy and effective systems. This study explores the impact of different human-AI collaboration modes on collaboration effectiveness and users' reliance on AI recommendations, particularly in the context of detecting stereotypical biases. Drawing on the literature of conjoined agency between humans and AI, this study differentiates between three distinct modes of human-AI collaboration— "human initiates task and AI assists," "AI screens all tasks and human assists," and "AI automation with human oversight"—and examines their varying effects on bias detection and human acceptance of AI recommendations. Additionally, we test the moderating role of equivocality in decision-making. Our study employs a 4x2 experimental design, including a fully manual control group, to test these hypotheses. This research contributes to the theoretical understanding of human-AI interaction and provides practical insights for designing more equitable and trusted AI systems.

Keywords Human-AI Collaboration, Stereotypical Bias Detection, Reliance on AI Recommendation, Equivocality

1 Introduction

As artificial intelligence (AI) becomes increasingly integrated into decision-making processes across domains such as healthcare, finance, and legal systems, concerns have arisen about the potential for these systems to reinforce and amplify existing human biases (Lai et al. 2023). Stereotypical biases, which are subtle yet deeply embedded in social structures, present significant challenges to ensuring fair and equitable outcomes in AI-driven decisions. While AI systems are praised for their ability to process vast amounts of data and identify patterns beyond human capabilities, they are not immune to the biases present in their training data, which can perpetuate historical and societal inequalities (Schemmer et al. 2022). Humans, on the other hand, also struggle with bias detection due to cognitive biases, errors in judgment, and the influence of social norms. Prior research highlights how these cognitive biases can lead to suboptimal decisions, particularly in complex or high-stakes scenarios (Kahneman et al. 2001). Moreover, while AI excels at data processing, it lacks the contextual understanding and ethical reasoning needed to fully address and mitigate biases (Buolamwini and Gebru 2018).

Given these challenges, it is crucial to explore how different human-AI collaboration modes can impact the effectiveness of bias detection and, importantly, how these modes influence human acceptance of AI recommendations. Prior research suggests that the success of human-AI collaboration in decision-making depends not only on the technical capabilities of the AI but also on the degree to which humans trust and accept AI's input (Dietvorst et al. 2015). This is in line with the broader research on trust in automation (Hoff et al 2015, Yang et al 2017). Factors such as transparency, perceived accuracy, and the ability to understand AI's reasoning play critical roles in building this trust. This study seeks to investigate how varying modes of human-AI collaboration affect human acceptance and confidence in AI recommendations, particularly in the context of detecting and mitigating stereotypical biases.

To address the above research gaps, this study investigates the impact of different human-AI collaboration modes on users' acceptance and confidence in AI recommendations, particularly in the context of detecting stereotypical biases. Drawing on the literature of conjoined agency between human and AI (Jarrahi 2018; Murray et al. 2021), this study differentiates between three distinct modes of human-AI collaboration— "human initiates task and AI assists," "AI screens all tasks and human assists," and "AI automation with human oversight"—and examines their varying effects on bias detection and human acceptance of AI recommendations. Additionally, we test the moderating role of equivocality in decision-making, which involves situations with multiple, often conflicting interpretations. Accordingly, we propose that the level of human control and involvement in the decision-making process will affect their acceptance and confidence in AI-generated outputs. Specifically, we suggest that modes where humans maintain a significant role in the decision-making process, such as "human initiates task and AI assists," are expected to lead to higher accuracy in decision-making. Additionally, these modes are anticipated to result in greater confidence in AI recommendations and a more effective response to equivocal decision-making tasks. In this research-in-progress paper, we employ a 4x2 experimental design, including a fully manual control group, to test these hypotheses and provide preliminary insights into optimizing human-AI collaborations for more equitable decision-making. By examining these dynamics, we aim to provide insights into optimizing human-AI interactions for more equitable and effective decision-making. This research will contribute to the growing body of literature on AI ethics and human-AI collaboration by offering a nuanced understanding of how to foster productive and trusting relationships between humans and AI in the context of bias detection.

2 Theoretical Foundation

2.1 Challenges in Stereotypical Bias Detection

Stereotypical bias is ingrained in many aspects of life, subtly influencing social interactions, decision-making processes, and systems, thus reinforcing stereotypes and perpetuating inequality (Ferrara 2023). Mehrabi et al. (2021) categorized bias in an AI system into three stages: "data to algorithm," occurring during model training; "algorithm to user," when recommendations are generated; and "user to data," involving the user's interpretation of recommendations influenced by their inherent biases. However, the algorithms utilized by AI are often driven by values such as efficiency, optimization, and objectivity, rather than fairness, leading to the potential for unnoticed undesirable effects (Tarafdar et al. 2023; Dolata et al. 2021).

Detecting and addressing this bias is essential for promoting fairness and inclusivity, ensuring that decision-making—whether human or AI-driven—does not perpetuate harmful biases. Such biases can skew perceptions and lead to unjust outcomes, particularly affecting marginalized groups in domains

like hiring (Upadhyay and Khandelwal 2018), legal judgments (Parikh et al. 2023), and education (AlAfnan et al. 2023; Eke 2023), and product design (Ostrowski et al. 2022; Das et al. 2023). As AI systems become more integrated into decision-making, addressing biases within these systems is crucial to prevent the amplification of existing human biases.

Humans often struggle to detect subtle stereotypical biases due to their own cognitive biases and judgment errors (Kahneman et al. 2001) and the interpersonal differences (Li & Holtta-Otto 2023). These challenges are exacerbated by factors like fatigue and attentional lapses, making consistent bias recognition difficult, especially in complex scenarios. While AI systems can process large datasets and identify patterns, they are not immune to biases, often reflecting the historical and societal biases present in their training data (Buolamwini and Gebru 2018). Moreover, AI systems lack the contextual understanding and ethical reasoning needed to effectively identify and mitigate biases, and their reliance on predefined metrics and lack of interpretability can hinder transparent decision-making. Given the limitations of both humans and AI in bias detection, this study explores the potential complementarity of human and AI capabilities, evaluates how different collaboration modes influence the ways in which humans interpret and adopt the recommendations provided by AI.

2.2 Conjoined Agency of Human and AI

The concept of *conjoined agency* between humans and AI reflects a paradigm shift in understanding decision-making processes as increasingly co-constructed through the interaction of human judgment and machine intelligence. Rather than operating independently, humans and AI are seen as interdependent agents, each contributing unique strengths. Jarrahi (2018) emphasizes the complementary roles of humans and AI, identifying key task attributes like equivocality that shape their interaction. *Equivocality*, where multiple, contrasting interpretations exist, particularly requires human intuition and ethical reasoning to complement AI's pattern recognition capabilities (Mittelstadt et al. 2016; Kahneman et al. 2021). Research highlights the essential interplay between human intuition and AI capabilities. Gigerenzer and Gaissmaier (2011) note that human heuristics are crucial when AI's algorithmic approaches fall short in complex scenarios. While AI excels at data processing, it often struggles with the nuanced interpretations required in ambiguous situations, making human insight critical for balanced decision-making (Gillespie et al. 2020; Lee et al. 2022; Murray et al. 2021).

The practical implications of conjoined agency are evident in AI-supported environments. Studies show that AI can handle routine tasks efficiently, allowing human experts to focus on complex cases requiring ethical judgment, as seen in medical fields like cancer screening (Hemmer et al. 2019; Topol 2019) or engineering design (Allison et al 2022). Additionally, when humans use AI tools to refine their judgments, it leads to more accurate outcomes, particularly in finance and healthcare, where human intuition and AI's analytical power synergize effectively (Green and Chen 2019; Caruana et al. 2015). Further, elevating AI into a more proactive collaborative role rather than only a tool has potential to enhance task performance by managing routine processes, freeing human decision-makers to focus on more complex and high-level decisions (Amershi et al. 2014, McComb et al 2023).

2.3 Human-AI Collaboration Modes

Building on these insights, this study focuses on three distinct collaboration modes that explore the complementarity of human and AI capabilities in decision-making. The first mode is "*human initiates task and AI assists*". In this mode, humans initiate tasks, but AI significantly aids the decision-making process by offering support through data analysis, pattern recognition, or predictive insights (Green and Chen 2019). This collaboration leverages AI's strengths while allowing humans to retain control over final decisions, which is particularly useful in areas where humans may be limited by cognitive biases or information overload (Kahneman and Tversky 1974).

The second mode is "*AI screens all tasks and human assists*". Here, AI takes the lead in screening and processing tasks, resolving those within its capability. It evaluates each task and assigns a confidence level to its response. If the confidence is high, the AI processes the task autonomously. For tasks where the confidence is lower, or when uncertainty arises, it delegates them to a human for final judgment. This mode is particularly relevant in contexts where AI efficiently handles routine tasks but requires human intervention for complex or ambiguous situations (Hemmer et al. 2019).

The third mode is "*AI automation with human oversight*," where AI handles the decision-making process from start to finish, but humans review and approve the final recommendations. In this mode, humans rely on AI-generated decisions, particularly in routine tasks, while ensuring alignment with ethical standards and addressing any potential biases (Buolamwini and Gebru 2018). Both this mode and the "AI screens all tasks and human assists" mode allow AI to process all tasks initially. However,

in the "AI screens all tasks and human assists" mode, only uncertain tasks are delegated to humans for final recommendations, whereas in this mode, humans review and approve all tasks before the final recommendation. This approach maximizes AI efficiency while maintaining necessary human oversight. However, it is also important to recognize that the quality of the recommendation outcome is contingent upon the individual's expertise in the relevant domain and their ability to identify biases, there are several challenges associated with human oversight, as noted by Dolata et al. (2021) and Sullivan et al. (2024). For instance, algorithms may influence human attention, individuals may misinterpret algorithmic outputs, or they may lack adequate information or a comprehensive understanding of the problem context.

2.4 Human Reliance on AI Recommendations

In real-life decision-making, even when AI significantly assists or automates tasks, humans often remain the final decision-makers. This oversight means the success of any human-AI collaboration hinges not only on AI's technical capabilities but also on the degree to which humans rely on its recommendations. Human reliance on AI is critical, especially in scenarios where AI is used to detect and mitigate bias (Lai et al. 2023). It is essential that humans rely on AI appropriately—not too little, risking missed opportunities for fairness, nor too much, risking the uncritical acceptance of biased AI outputs. Human acceptance and confidence in AI recommendations are thus crucial. If humans distrust AI, they may disregard its recommendations, leading to suboptimal decisions (Dietvorst et al. 2015). Conversely, over-reliance without sufficient understanding can also result in poor outcomes. Therefore, understanding how reliance varies across collaboration modes is essential for optimizing human-AI interactions and ensuring more accurate, fair, and effective decisions.

Research shows that human reliance on AI is significantly influenced by the transparency and perceived accuracy of AI recommendations, and the trust users have in the AI's decision-making process. When AI lacks transparency or fails to clearly convey its reasoning, users are more likely to feel uncertain, leading to reduced reliance (Gustafsson et al. 2020; Sundar 2020). The opacity of AI's decision-making can undermine confidence, as users may struggle to understand or trust its conclusions (Jacovi et al. 2021; McKnight et al. 2021). Confidence in AI is strongly linked to its demonstrated accuracy and reliability over time (Logg et al. 2019; Grote and Berens 2020, Yang et al 2017). Consistent accuracy builds trust and reliance, but this can vary significantly across collaboration modes, especially in tasks with high equivocality (Papenmeier et al. 2019; Lee 2018). These dynamics highlight the need to examine how reliance on AI varies to ensure that human-AI partnerships lead to better decision-making outcomes.

3 Hypothesis Development

3.1 Effects of Human-AI Collaboration Modes

The detection of stereotypical biases is a complex challenge that requires the complementary strengths of both human intuition and AI's data-processing capabilities. Stereotypical biases are often subtle and deeply embedded in societal norms, making them difficult for either humans or AI to detect independently. Humans possess the ethical reasoning and contextual understanding needed for understanding the data. For example, while large language models have been developed to specifically understand natural language including context, comparing human and natural language processing understanding reveals how humans infer context much better than a large language model (Fataliyev et al 2023). This contextual understanding and the ability consciously acknowledge biases or differences in values via frameworks such design justice framework (Das et al 2032) can help humans detect or counteract biases. However, cognitive biases such as confirmation bias and anchoring or interpersonal differences can impair their ability to objectively detect these biases (Kahneman and Tversky 1974, Li & Holtta-Otto 2023). Conversely, AI systems, while proficient at processing large datasets and identifying patterns, are prone to inheriting and perpetuating the biases present in their training data (Buolamwini and Gebru 2018). A collaborative approach that integrates human judgment with AI's analytical strengths is therefore expected to be more effective in detecting stereotypical biases, as it leverages the unique capabilities of both humans and AI. While we acknowledge the full fairness of any system can depend on the perspective (Dolata et al 2021, Sullivan et al 2024), this synergy allows for a more thorough examination of potential biases, leading to more accurate and equitable outcomes compared to AI-driven decision-making with human oversight.

H1: *Human-AI collaboration modes (i.e., "human initiates task and AI assists" and "AI screens all tasks and human assists") will be more effective in detecting stereotypical biases compared to AI-driven decision-making with human oversight (i.e., "AI automation with human oversight").*

In scenarios where humans initiate tasks and leverage AI to refine and enhance their judgments, the combination is likely to produce more accurate and reliable decisions. This is particularly true in complex situations where human intuition and ethical considerations play a crucial role. The "human initiates task and AI assists" mode is characterized by greater human control over the decision-making process, with AI providing support through data analysis, pattern recognition, or predictive insights. This mode allows humans to leverage AI's strengths while maintaining final decision-making authority, which is particularly beneficial in complex scenarios requiring ethical judgment. Prior studies suggest that this combination of human initiation and AI support is likely to produce more accurate and reliable decisions (Green and Chen 2019). In contrast, the "AI screens all tasks and human assists" mode, where AI takes the lead and humans intervene only as needed, may not provide the same level of accuracy in bias detection, as it relies more heavily on AI's outputs.

H2: *The "human initiates task and AI assists" mode will lead to higher accuracy in decision-making and bias detection compared to the "AI screens all tasks and human assists" mode.*

Reliance on AI recommendations is crucial for the success of human-AI collaboration, with trust heavily influenced by the degree of control and transparency in the decision-making process (Wang et al. 2019, Yang 2017). The literature on trust in automation emphasizes that users are more likely to trust and rely on AI when they feel a sense of agency, where they can directly influence decisions (Lee and See 2004), but this is impacted by cognitive biases such s over condidence. In the "human initiates task and AI assists" mode, humans maintain primary control over the decision-making process, which allows them to leverage AI's analytical capabilities while retaining final authority over decisions. This sense of control is critical, as it enables users to validate AI suggestions and feel more confident in the outcomes. Research by Dzindolet et al. (2003) supports this, showing that when individuals perceive themselves as having control and can scrutinize AI inputs, their trust in AI systems increases, leading to greater acceptance of AI-generated recommendations.

On the other hand, in the "AI screens all tasks and human assists" mode, where AI leads the decision-making process and humans intervene only as needed, users may experience diminished confidence in AI outputs. This is because the AI's reasoning for bias detection is not fully transparent or understandable to the user, which can lead to uncertainty and hesitation in relying on the AI's decisions. The lack of transparency of AI's screening process, combined with reduced human agency can undermine user reliance (Gustafsson et al. 2020). As a result, users may be less likely to fully rely on AI recommendations in this mode.

H3: *"Human initiates task and AI assists" mode will lead to higher human reliance on AI recommendations compared to the "AI screens all tasks and human assists" mode.*

3.2 Effects of Task Equivocality

The detection of stereotypical biases becomes increasingly complex in tasks characterized by high equivocality—situations where multiple, often conflicting interpretations exist, and where the correct course of action is not immediately clear. In such contexts, the nuanced understanding and ethical reasoning that humans bring to the decision-making process are crucial (Jarrahi 2018; Murray et al. 2021). The "human initiates task and AI assists" mode allows humans to leverage their contextual knowledge and ethical considerations while receiving support from AI in processing large datasets and identifying patterns. This collaboration is particularly effective in high-equivocality scenarios, where human judgment plays a critical role in interpreting ambiguous data and ensuring that biases are accurately detected and addressed (Mittelstadt et al. 2016). Conversely, in the "AI screens all tasks and human assists" mode, the AI's lead role may result in less effective bias detection, especially in complex, ambiguous tasks where human intervention is limited to post-hoc corrections (Gillespie et al. 2020). Therefore, the presence of equivocality is expected to amplify the effectiveness of the "human initiates task and AI assists" mode in bias detection.

H4: *The presence of equivocality in decision-making tasks will moderate the relationship between collaboration mode and bias detection effectiveness, such that the effect of the "human initiates task and AI assists" mode over the "AI screens all tasks and human assists" mode will be more pronounced in tasks characterized by high equivocality.*

Equivocality in decision-making tasks often leads to uncertainty, making it difficult for AI systems to provide clear recommendations without human input. In these situations, humans are more likely to rely on AI when they maintain a significant role in the decision-making process, as in the "human initiates task and AI assists" mode. This mode facilitates dynamic interaction, where AI analysis informs human judgment, thereby increasing confidence in AI's recommendations. When tasks are highly equivocal, human interpretation becomes critical, reinforcing reliance on AI that is supported by human

oversight. Conversely, in the "AI screens all tasks and human assists" mode, where AI leads and human intervention is minimal, the lack of transparency and reduced human agency can diminish reliance on AI, particularly in equivocal scenarios (Gustafsson et al. 2020; Jacovi et al. 2021). Thus, equivocality is expected to enhance reliance on AI in the "human initiates task and AI assists" mode compared to the "AI screens all tasks and human assists' mode".

H5*: The presence of equivocality in decision-making tasks will moderate the relationship between collaboration mode and reliance on AI recommendations, such that the effect of the "human initiates task and AI assists" mode over the "AI screens all tasks and human assists" mode will be more pronounced in tasks characterized by high equivocality.*

4 Research Method

4.1 Experimental Design

To investigate our hypotheses regarding the influence of human-AI collaboration modes on bias detection and human acceptance of AI recommendations, we adopt a 4 (collaboration mode: "human manual" vs. "human initiates task and AI assists" vs. "AI screens all tasks and human assists" vs. "AI automation with human oversight") × 2 (task equivocality: low equivocality vs. high equivocality) between-subjects design with random assignment. Specifically, in the "human manual" condition, participants will complete the task independently without any AI assistance. In the "human initiates task and AI assists" condition, participants will first assess a scenario and then receive AI-generated recommendations before making a final decision. In the "AI screens all tasks and human Assists" condition, the AI first screens all the scenarios, flagging cases with high uncertainty or potential bias for the participant to review and decide upon. In this case, only a subset of the scenarios is presented to the participants, who are informed that these are the scenarios the AI is unsure about. The AI's recommendations are shown to the participants, even though they carry a high level of uncertainty. In the "AI automation with human oversight" condition, the AI will autonomously assess the scenario and provide a recommendation, which the participant can then review. The participant's role will be to either accept the AI's decision or intervene if they believe the AI has made an error, ensuring human oversight in the decision-making process.

To differentiate between low and high equivocality of bias detection tasks, we adopted tasks from prior studies conducted by Lin et al. (2023) and Nadeem et al. (2021). For low equivocality tasks, scenarios are designed with clarity, where the presence or absence of stereotypical bias is straightforward and unambiguous, such as a job candidate described with explicit gender-based language (e.g., "She is highly organized, a typical trait for women in administrative roles"). In contrast, high equivocality tasks involve scenarios where the presence of bias is more ambiguous and open to interpretation, with subtle or mixed cues that make it challenging to determine if bias is present (e.g., "She is known for her nurturing approach in managing the team").

4.2 Experimental Procedure and Measurements

As part of a larger human-AI collaboration study, participants will be recruited from Prolific for a 15-minute online experiment, presented as assisting in the evaluation of an AI tool, and will be compensated with US$5. Four inclusion criteria will be used: participants must be located in the United States, fluent in English, not among the top 4% of workers in terms of survey volume (to avoid professional survey-takers) and have a worker approval rate above 98% to ensure high-quality responses. Participants who fail attention checks (e.g., comprehension questions related to the AI's function or key characteristics) will be excluded from the analysis. Before the experiment, the implicit association test (IAT) will be employed to assess participants' implicit biases, revealing unconscious stereotypical preferences through rapid sorting tasks (Greenwald et al. 1998). Then participants will be randomly assigned to one of the experimental conditions. The experiment includes six stereotype detection tasks. Each task involves an evaluation of a scenario for stereotypical bias rating on a 7-point scale based on a paragraph of text.

The effectiveness of detecting stereotypical bias is assessed using two metrics: the accuracy of the final decision-making outcome with AI's assistance and the impact of AI's advice on human decision-making. The first dependent variable, the rating accuracy, is calculated by comparing the difference between participants' revised ratings and the pre-validated true score using the dataset validation process outlined by Nadeem et al. (2021). The second metric, based on the methodology of Logg et al. (2019), calculates the Weight of Advice (WoA). This dependent variable is defined as the ratio of the difference

between the initial and revised ratings to the difference between the initial rating and the AI's recommendation. The formula is given by:

$$Weight\ of\ Advice = \frac{I-R}{I-A}$$, Where:

- I = participant's initial rating to the scenario
- R = participant's revised rating to the scenario after receive AI recommendation
- A = AI's recommend rating

After completing the task, participants were asked to complete a survey. We first asked participants to indicate their perceived fairness (Franke et al., 2013). We also asked participants about their trusting intentions (McKnight et al. 2002) by inquiring if they would recommend using the AI systems. Afterward, we collected demographic information (age, gender and education level), familiarity with the task (adapted from Gefen, 2000), familiarity with AI (Logg et al. 2019) and disposition to trust (Gefen, 2000) as control variables. All items are measured on a 7-point Likert scale.

5 Conclusion

Designing human-AI collaboration modes that ensure effective and fair decision-making is a crucial challenge at the intersection of human-AI interaction and organizational management. Drawing from literature on conjoined agency between humans and AI, this research-in-progress paper investigates how distinctive collaboration modes impact decision-making effectiveness and users' reliance on AI recommendations during stereotype detection tasks. This study offers three key contributions: First, it empirically shows how collaboration design shapes user reliance on AI, especially in ethically sensitive contexts. Second, it underscores the importance of human judgment and task context in determining the effectiveness of AI-driven decisions. Third, it explores the moderating effect of equivocality in decision-making, demonstrating how task ambiguity influences the effectiveness of collaboration modes. This research provides practical insights for AI designers on creating systems that balance AI's analytical strengths with essential human oversight, promoting more equitable and trusted AI systems.

6 References

AlAfnan, M. A., Dishari, S., Jovic, M., and Lomidze, K. 2023. "ChatGPT as an Educational Tool: Opportunities, Challenges, and Recommendations for Communication, Business Writing, and Composition Courses." *Journal of Artificial Intelligence and Technology* (3:2), pp. 60-68.

Allison, J. T., Cardin, M. A., McComb, C., Ren, M. Y., Selva, D., Tucker, C., ... & Zhao, Y. F. (Eds.). (2022). Artificial intelligence and engineering design. *Journal of mechanical design*, *144*(2), 020301.

Amershi, S., Cakmak, M., Knox, W.B., and Kulesza, T. (2014). "Power to the People: The Role of Humans in Interactive Machine Learning." *AI Magazine* (35:4), pp. 105-120.

Buolamwini, J., and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." *Proceedings of the Conference on Fairness, Accountability, and Transparency* (FAT), pp. 77-91.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. 2015. "Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission," *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 1721-1730.

Das, M., Roeder, G., Ostrowski, A. K., Yang, M. C., & Verma, A. (2023). What Do We Mean When We Write About Ethics, Equity, and Justice in Engineering Design?. *Journal of Mechanical Design*, *145*(6), 061402.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). "Algorithm aversion: people erroneously avoid algorithms after seeing them err." *Journal of Experimental Psychology: General* (144:1), pp. 114.

Dolata, M., Feuerriegel, S., & Schwabe, G. (2022). "A sociotechnical view of algorithmic fairness." *Information Systems Journal*, 32(4), 754-818.

Dzindolet, M.T., Peterson, S.A., Pomranky, R.A., Pierce, L.G., and Beck, H.P. (2003). "The Role of Trust in Automation Reliance." *International Journal of Human-Computer Studies* (58:6), pp. 697-718.

Eke, D. O. 2023. "ChatGPT and the Rise of Generative AI: Threat to Academic Integrity?" *Journal of Responsible Technology* (13), p. 100060.

Ferrara, E. 2023. "Should ChatGPT Be Biased? Challenges and Risks of Bias in Large Language Models." *arXiv preprint* (arXiv:2304.03738).

Fataliyev, S. N., Beck, D., & Holtta-Otto, K. (2023, November). Predicting Empathic Accuracy from User-Designer Interviews. In *Proceedings of the 21st Annual Workshop of the Australasian Language Technology Association* (pp. 125-129).

Franke, G.R., Crown, D.F., and Spake, D.F. (2013). "A Critical Review of the Effects of Product Stereotypes on Advertising Evaluations: The Role of Implicit and Explicit Biases." *Journal of Advertising* (42:3-4), pp. 158-169.

Gefen, D. (2000). "E-commerce: The Role of Familiarity and Trust." *Omega* (28:6), pp. 725-737.

Gigerenzer, G., and Gaissmaier, W. (2011). "Heuristic Decision Making." *Annual Review of Psychology* (62), pp. 451-482.

Gillespie, T., Boczkowski, P. J., and Foot, K. A. 2020. *Media Technologies: Essays on Communication, Materiality, and Society*. Cambridge, MA: MIT Press.

Green, P., and Chen, K. (2019). "Human-AI Collaboration in Decision-Making: The Role of Human Judgment in Complex Scenarios." *Journal of Behavioral Decision Making* (32:3), pp. 456-468.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. 1998. "Measuring Individual Differences in Implicit Cognition: The Implicit Association Test." *Journal of Personality and Social Psychology* (74:6), p. 1464.

Grote, G., and Berens, S. 2020. "On the Role of Context in Decision-Making: An Integrative Approach," *Journal of Organizational Behavior* (41:7), pp. 663-678.

Gustafsson, M., Binns, R., and Jirotka, M. (2020). "Ethical AI: Understanding Human Agency and the Impact of Automation." *Philosophy & Technology* (33:4), pp. 629-650.

Hemmer, P., Steyvers, M., and Lee, M.D. (2019). "Collaborative Decision-Making in the Face of Uncertainty: The Role of Confidence and Explanation." *Cognitive Science* (43:2), pp. 1-21.

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human factors*, *57*(3), 407-434.

Jacovi, A., Marasović, A., and Goldberg, Y. (2021). "Aligning AI Trustworthiness with Human Trust." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics* (ACL), pp. 2606-2617.

Jarrahi, M.H. (2018). "Artificial Intelligence and the Future of Work: Human-AI Symbiosis in Organizational Decision Making." *Business Horizons* (61:4), pp. 577-586.

Kahneman, D., and Tversky, A. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science* (185:4157), pp. 1124-1131.

Kahneman, D., Sibony, O., and Sunstein, C. R. 2021. *Noise: A Flaw in Human Judgment*. New York: Little, Brown Spark.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q. V., and Tan, C. (2023). "Towards a Science of Human-AI Decision Making: An Overview of Design Space in Empirical Human-Subject Studies." *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1369-1385.

Lee, J. D. 2018. "Trust, Control, and Cognitive Automation in the Use of Decision Support Systems," *Human Factors* (60:2), pp. 187-202.

Lee, J.D., and See, K.A. (2004). "Trust in Automation: Designing for Appropriate Reliance." *Human Factors* (46:1), pp. 50-80.

Lee, M. K., Kusbit, D., Metsky, E., and Dabbish, L. 2022. "Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers," *Proceedings of the 2022 ACM Conference on Human Factors in Computing Systems*, pp. 1603-1612.

Li, J., & Hölttä-Otto, K. (2023). Inconstant Empathy—Interpersonal Factors That Influence the Incompleteness of User Understanding. *Journal of Mechanical Design*, *145*(2), 021403.

Lin, C.-C., Akuhata-Huntington, Z., and Hsu, C.-W. 2023. "Comparing ChatGPT's Ability to Rate the Degree of Stereotypes and the Consistency of Stereotype Attribution with Those of Medical Students in New Zealand in Developing a Similarity Rating Test: A Methodological Study." *Journal of Educational Evaluation for Health Professions* (20:17).

Logg, J.M., Minson, J.A., and Moore, D.A. (2019). "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment." *Organizational Behavior and Human Decision Processes* (151), pp. 90-103.

McComb, C., Boatwright, P., & Cagan, J. (2023). Focus and modality: defining a roadmap to future AI-human teaming in design. *Proceedings of the Design Society*, *3*, 1905-1914.

McKnight, D.H., Carter, M., and Burgoon, J.K. (2021). "Trust in Technology: Development of a Multidimensional Trust in Technology Scale." *MIS Quarterly* (45:1), pp. 265-296.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). "A survey on bias and fairness in machine learning." *ACM computing surveys* (CSUR), 54(6), 1-35.

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., and Floridi, L. 2016. "The Ethics of Algorithms: Mapping the Debate," *Big Data & Society* (3:2), pp. 1-21.

Murray, A., Rhymer, J. E. N., and Sirmon, D. G. (2021). "Humans and Technology: Forms of Conjoined Agency in Organizations," *Academy of Management Review* (46:3), pp. 552-571.

Nadeem, M., Bethke, A., and Reddy, S. 2021. "StereoSet: Measuring Stereotypical Bias in Pretrained Language Models." *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356-5371.

Ostrowski, A. K., Walker, R., Das, M., Yang, M., Breazea, C., Park, H. W., & Verma, A. (2022, August). Ethics, equity, & justice in human-robot interaction: A review and future directions. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)* (pp. 969-976). IEEE.

Papenmeier, A., Englebienne, G., and Theune, M. (2019). "Managing Uncertainty in Human-AI Interaction: A Study on the Effects of AI Transparency and Confidence." *International Journal of Human-Computer Studies* (127), pp. 121-132.

Schemmer, M., Hemmer, P., Kühl, N., Benz, C., & Satzger, G. (2022). "Should I Follow AI-based Advice? Measuring Appropriate Reliance in Human-AI Decision-making." *arXiv preprint* (arXiv:2204.06916).

Sullivan, R., Veen, A., & Riemer, K. (2024). "Furthering engaged algorithmic management research: Surfacing foundational positions through a hermeneutic literature analysis." *Information and Organization*, 34(4), 100528.

Sundar, S.S. (2020). "The Value of Machine Heuristics: How Perceptions of Automation Shape Trust in AI." *Journal of Computer-Mediated Communication* (25:4), pp. 292-307.

Tarafdar, M., Page, X., & Marabelli, M. (2023). "Algorithms as co-workers: Human algorithm role interactions in algorithmic work." *Information Systems Journal,* 33(2), 232-267.

Topol, E. J. 2019. *Deep Medicine: How Artificial Intelligence Can Make Healthcare Human Again*. New York: Basic Books.

Upadhyay, A. K., and Khandelwal, K. 2018. "Applying Artificial Intelligence: Implications for Recruitment." *Strategic Human Resource Review* (17:5), pp. 255-258.

Wang, W., Lee, H., and Stolterman, E. (2019). "Trust in Human-AI Collaboration: Designing for Appropriate Trust in Artificial Intelligence." *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI), pp. 1-12.

Yang, X. J., Unhelkar, V. V., Li, K., & Shah, J. A. (2017, March). Evaluating effects of user experience and system transparency on trust in automation. In *Proceedings of the 2017 ACM/IEEE international conference on human-robot interaction* (pp. 408-416).

Copyright