Trust and Reliance in Evolving Human-AI Workflows (TREW)

Zahra Ashktorab*
zashktorab@gmail.com
IBM Research
New York, USA

Gagan Bansal*
gaganbansal@microsoft.com
Microsoft Research
Seattle, USA

Zana Buçinca*
zanabucinca@g.harvard.edu
Harvard University
USA

Kenneth Holstein*
kjholste@cs.cmu.edu
Carnegie Mellon University
USA

Jessica Hullman*
jhullman@northwestern.edu
Northwestern University
USA

Alison Smith-Renner*
arenner@dataminr.com
Dataminr
New York, USA

Tongshuang Wu*
sherryw@cs.cmu.edu
Carnegie Mellon University
Pittsburgh, USA

Wenjuan Zhang*
wzhang@dataminr.com
Dataminr
New York, USA

ABSTRACT

State-of-the-art AIs, including Large Language Models (LLMs) like GPT-4, now possess capabilities once unique to humans, such as coding, idea generation, and planning. Advanced AIs are now integrated into a plethora of platforms and tools, including GitHub Copilot, Bing Chat, Bard, ChatGPT, and Advanced Data Analytics. In contrast to conventional, specialized AIs that typically offer singular solutions, these LLMs redefine human-AI dynamics, with a growing trend toward humans viewing them as collaborative counterparts. This shift leads to enhanced dialogues, negotiations, and task delegation between humans and AI. With these rapid advancements, the nature of human roles in the AI collaboration spectrum is evolving. While our previous workshops CHI TRAIT 2022 and 2023 delved into the trust and reliance concerning traditional AIs, the pressing question now is: how should we measure trust and reliance with these emerging AI technologies? As these systems witness widespread adoption, there's also a need to assess their impact on human skill development. Does AI assistance amplify human skill progression, or does it inadvertently inhibit it? Considering the multifaceted challenges and solutions that revolve around human-AI interactions, we invite experts from diverse fields, including HCI, AI, ML, psychology, and social science. Our aim is to bridge communication gaps and facilitate rich collaborations across these domains.

CCS CONCEPTS

• **Human-centered computing → Empirical studies in HCI**; *Interactive systems and tools*; • **Computing methodologies →** Machine learning.

KEYWORDS

trust, human-centered artificial intelligence, reliance, uncertainty

ACM Reference Format:
Zahra Ashktorab, Gagan Bansal, Zana Buçinca, Kenneth Holstein, Jessica Hullman, Alison Smith-Renner, Tongshuang Wu, and Wenjuan Zhang. 2024. Trust and Reliance in Evolving Human-AI Workflows (TREW). In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24), May 11–16, 2024, Honolulu, HI, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3613905.3636319

1 INTRODUCTION AND BACKGROUND

One recent significant development in AI has been the development of LLMs (e.g., GPT-4, Bard, Claude, etc). Numerous studies have demonstrated that these models now possess the ability to perform a wide range of tasks that were once considered exclusive to people [2]. For instance, LLMs can generate annotations more effectively than crowdworkers [8], write advanced programs efficiently [24], and achieve impressive results in academic and professional exams [20]. These observations suggest that LLMs can have significant social and economic implications, potentially reshaping the workforce by supplementing and even driving the creation of new human job and roles [5]. However, there are also widespread concerns about the safety and reliability of deploying LLMs. These models often exhibit unpredictable performance [7] and instability [12], making it challenging to anticipate the consequences of their development and deployment. The tension between LLM competence and safety concerns underscores the importance of a future that emphasizes human-AI collaboration [1, 10, 11, 19] — to establish a coexistence, where AIs handle tasks within their capabilities while humans handle the rest.

While the topic of "human-AI collaboration" has been studied for decades, it is now undergoing a transformation because of the emergence of LLMs. In particular, as these models become more versatile, humans shift from viewing AI systems as single-purpose supportive roles that provide suggestions on dedicated tasks, to more *collaborative peers* that can react more freely to various types of inquiries and can take initiative beyond reacting to human requests. As such, it is essential to understand **how the role**

[*] Equal contribution from all authors. Names ordered alphabetically.

CHI EA '24, May 11–16, 2024, Honolulu, HI, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0331-7/24/05
https://doi.org/10.1145/3613905.3636319

played by the AI versus the user change with these advances, and how to best design for effective workflows. One example of such role shift is already visible in the context of programming: GitHub Copilot (a commercialized programming assistant) has been reported to contribute up to 40% of code to programmers' code bases [26] and, as a result, programmers now delegate code writing to AIs, and switch their focus to code verification [18]. Beyond programming assistants, these AIs are also increasingly considered as potential pair programmers [17], which further testifies to how humans may view LLMs closer to complex collaborators than tools. Abstracting and defining such shifts in human and AI roles will directly impact how we contextualize the future work on human-AI collaboration.

In particular, the topic of trust and reliance in human-AI collaboration deserves significant attention. Unlike typical discriminative models, LLMs, influenced by vast language patterns, can generate seemingly plausible, yet non-factual statements (i.e., "hallucinations"). For example, a lawyer was sanctioned for referencing a ChatGPT-generated, hallucinated case [23]. Such incidents underscore the importance of rethinking trust and reliance in human-AI interaction in the context of LLMs. To ensure that human-AI collaborations do more good than harm, it is vital that we understand, measure, and shape human-AI trust and reliance; for example, **when is user trust and reliance on AI warranted or appropriate, how does it evolve over the course of (long-term) human-AI collaborations**, which factors (individual, contextual, etc) influence it, and how can we design interventions that guard against harms caused by inappropriate trust or reliance and instead promote more appropriate reliance within human-AI collaborations? Indeed, recent years have seen renewed interest in human-AI trust within the AI and HCI community [3, 13, 15, 16, 22, 25] motivated by directions and approaches introduced in prior literature on human-automation interaction and human factors [14]. However, it is not yet well understood what generalizes from these works and what new considerations modern human-AI collaborations introduce [4, 9].

While trust and reliance continue to be essential for ensuring the appropriate usage of AI systems, the rapid transformation of people's workflows by LLM-powered systems has raised profound questions regarding the long-term impact on individuals receiving AI assistance. As humans constantly collaborate with the same AI on multiple tasks and through a longer time period (v.s. in traditional AIs either the task is fixed or the interaction is short), both the human and the AI learn and adapt to each other. On the one hand, it is not yet clear how the human collaborator evolves during the interaction with the AI system and whether the AI assistance helps or hinders their learning about the task (e.g., coding). Some recent evidence suggests that the current design of AI assistance, in which people are directly "handed the answer (i.e., the AI recommendation)" may hinder long-term human learning and skill development on the task they receive assistance with [6]. On the other hand, LLMs also adapt as they collect more human inputs, and certain explicit or implicit human feedback (or even feedback from the AI themselves) may have an impact on the capabilities of AIs [21]. The co-adaptation means the entities involved in collaborations, and thus the nature of collaboration, will be dynamic, and **capturing this changing factor enables us to look into longer term effects of human-AI collaboration**.

In sum, our goal is to revisit the topic of trust and reliance in in the evolving context of human-AI collaboration in the age of LLMs. Specifically, in this new context, we want to synthesize and conceptualize:

1. **how the AI roles change, and how the user roles have changed with these new AI**
2. **whether and how trust and reliance definitions and measures should be changed**
3. **the impact of these new interactions on users, especially their individual skills and learning**

Answering these questions is crucial to design these systems for human learning and skill improvement in addition to accuracy when completing AI-assisted tasks and fostering a more meaningful human-AI collaboration. As the problems and solutions involving AI and people are inter-disciplinary in nature, we will invite people with expertise in HCI, AI, ML, psychology, social sciences, or other relevant fields to foster closer communication and collaborations between multiple communities.

2 ORGANIZERS

Zahra Ashktorab is a Research Scientist at IBM Research, New York. Her research lies at the intersection of HCI and AI, primarily focusing on human-AI collaboration, with a special emphasis on enhancing the efficacy of these interactions.

Gagan Bansal is a a Senior Researcher at Microsoft Research, Redmond. He received his Ph.D. degree from the Allen School of Computer Science and Engineering at the University of Washington, Seattle. He conducts interdisciplinary research on Artificial Intelligence and Human-Computer Interaction with focus on developing human-centered AI systems for augmenting people.

Zana Buçinca is a Ph.D. Candidate at Harvard University. Her research lies at the intersection of Human-Computer Interaction and Artificial Intelligence. Informed by cognitive science theories, Zana designs, builds, and evaluates AI for decision-making support.

Kenneth Holstein is an Assistant Professor in Human-Computer Interaction at Carnegie Mellon University. His research interests lie at the intersection of HCI, AI, design, and cognitive science, focusing on the design, development, and evaluation of human-AI collaborative systems in complex social contexts.

Jessica Hullman is an Associate Professor of Computer Science at Northwestern University. Her research looks at how to design, evaluate, coordinate, and think about visual representations of data and model predictions for inference, decision making, and communication, including the effects of visualizing uncertainty on belief updating and potential for behaviorally induced feedback loops in visualizing model predictions in strategic settings.

Alison Smith-Renner is a Research Manager at Dataminr. Her research interests lie at the intersection of NLP and HCI, focusing on transparency and control for interactive NLP systems to engender appropriate trust and improve human performance. Alison received her Ph.D. from the University of Maryland, College Park. She has organized various workshops and tutorials on explainable AI, human-AI trust, and human-centered AI, including at IUI, CHI, NAACL, and TEI, and she has held senior committee roles at IUI, CHI, and EMNLP.

Sherry Tongshuang Wu is an Assistant Professor in the Human-Computer Interaction Institute at Carnegie Mellon University. Her research lies at the intersection of Human-Computer Interaction and Natural Language Processing, aiming to design, evaluate, build, and interact with AI systems that are compatible with actual human goals. Before joining CMU, Sherry received her Ph.D. degree from the University of Washington.

Isabel Zhang is a Staff Research Scientist at Dataminr. Her research interest lies in human interaction with various automation technologies, including Artificial Intelligence. Her research has focused on understanding human workflow, situational awareness, and overall system performance in collaborative human-automation workflows. Isabel received her Ph.D. in Human-Systems Engineering from North Carolina State University.

3 WEBSITE

We will reuse and update our website at https://chi-trew.github.io to advertise and disseminate the workshop's information and call for proposals. We will also use this website to share workshop contributions, including accepted papers, and support future engagement.

4 PRE-WORKSHOP PLAN

In addition to the website, we will advertise the workshop through email distribution lists at relevant conferences and research institutions (including, but not limited to, FAccT, IUI, CHI, ACL, and CSCW mailing lists), direct communication with colleagues in the field, and social media.

We will have a program committee (PC) with experts from diverse research organizations and backgrounds who will help us to curate the workshop by disseminating the call for papers and reviewing submissions. Last year, we had commitments from 33 PC members with expertise in human-AI interaction and related topics. These PC members contributed timely and thoughtful paper reviews and were crucial to the workshop success. We plan to re-invite many (if not all) of them this year, including:

Ben Shneiderman (University of Maryland), Hal Daume III (University of Maryland), Michael Bernstein (Stanford University), Krzysztof Gajos (Harvard University), Elena Glassman (Harvard University), Maria De-Arteaga (UT Austin), Alon Jacovi (Bar Ilan University), Matthew Kay (Northwestern University), Michael Terry (Google Research), Fan Du (Adobe Research), Victor Dibia (Microsoft Research), Vera Liao (Microsoft Research), Jenn Wortman Vaughan (Microsoft Research), Tim Miller (University of Melbourne), Jim Chen (University of Washington), Erin Chiou (ASU Adapt Lab), Ian Covert (University of Washington), Shi Feng (University of Maryland), Ella Glikson (Bar Ilan University), Maia Jacobs (Northwestern University), Joseph Janizek (University of Washington), Retno Larasati (The Open University), Brian Lim (National University of Singapore), Ishan Nigam (UT Austin), Marissa Radensky (University of Washington), Gonzalo Ramos (Microsoft Research), Jakob Schoeffer (Karlsruhe Institute of Technology), Tom Williams (Colorado School of Mines), Ming Yin (Purdue University), Tony Zhang (Fortiss), and Joyce Zhou (Cornell University).

Participants interested in giving a presentation at the workshop will need to submit a short paper (2-6 pages). Submission types will include, but are not limited to, position papers summarizing authors' existing research in the area and how it relates to the workshop theme, papers that review the related literature and offer a new perspective, and papers that describe work-in-progress research projects. We will encourage submissions that present diverse viewpoints on the workshop topics, and encourage participation across relevant fields, such as AI, HCI, and cognitive psychology. We will also create and advertise an industry track that encourages practitioners (who don't usually come to CHI) to submit papers that offer an industrial perspective or real-world approach to the workshop theme. We will use Easychair to collect and review these submissions. Each submission will be reviewed by at least two PC members and one organizing committee member. To accommodate as many participants as possible, we plan to maintain a 50%-60% acceptance rate (similar to last year), such that paper authors can get a chance to join our workshop. Among them, we plan to include 8-10 spotlight that will be presented for during paper sessions. The remaining papers will be presented as posters (onsite) and as 1-minute pre-recorded presentations (online) during coffee breaks (see the next section for details on the workshop schedule.)

As workshop organizers, we value and are committed to diversity, equity, and inclusion. We welcome and encourage the participation of people who identify with any historically marginalized or underrepresented group. Further, the listed platforms and technologies should not be a barrier to the participation of anyone interested in this workshop. If the technologies listed do not accommodate participants needs, we will work with participants to find alternative solutions.

5 WORKSHOP STRUCTURE

In order to facilitate more in-depth conversations, we have tailored this workshop for a group of 40-50 participants. If interested participants exceed this number after the initial advertisement of the workshop, we may adjust the workshop structure to accommodate a slightly higher number of participants. We plan to organize our proposal as a single-day workshop, from 9:00 AM to 5:00 PM local time (including breaks), in a hybrid format. We hope most participants join the in-person workshop but will also plan for synchronous online access to the workshop. We intend to use a combination of Zoom (for synchronized talks) and Slack (for virtual and asynchronous question-answering and online discussions). In our post-workshop survey, we found that participants were particularly impressed by the Slack space setup, which we intend to keep: we will have channels for workshop sessions, and threads for each accepted paper so participants can submit their targeted questions. We will strongly encourage keynote speakers, paper authors, and panelists to actively monitor and respond accordingly. The organizers will also work with the technical team at CHI 2024 to utilize provided streaming methods (with captioning for accessibility), so as to minimize the jump across platforms.

The tentative workshop schedule is detailed in Table 1. Since one of our goals is to synthesize knowledge and expertise from various communities and spur impactful future research, the workshop will dedicate sufficient time for group discussions and activities

Slot	Theme
09:00 – 09:15 (15min)	Welcome
09:15 – 10:15 (60min)	Keynote talk, potentially by a leading expert on human-AI collaboration
10:15 – 10:45 (30min)	Paper sessions 1
10:15 – 10:45 (30min)	Coffee break (concurrent with poster presentations)
10:45 – 11:30 (45min)	Panel with experts that have diverse and well-balanced expertise (Michael Bernstein, Stephanie Bell, Su Lin Blodgett, Jina Suh)
11:30 – 12:00 (30min)	Paper sessions 2
12:00 – 13:00 (60min)	Lunch break
13:00 – 14:30 (90min)	Group activity 1 *(60 min discussion + 30 min group result sharing)*
14:30 – 15:00 (30min)	Coffee break (concurrent with poster presentations)
15:00 – 16:30 (90min)	Group activity 2 *(60 min discussion + 30 min group result sharing)*
16:30 – 16:45 (15min)	Closing remarks

Table 1: Tentative schedule for the proposed single day workshop. The workshop will dedicate sufficient time for group discussions and activities (afternoon session) in addition to a knowledge-sharing and discussion in the form of a keynote, paper presentations, and an expert panel discussion (morning session).

(afternoon session) in addition to a knowledge-sharing and discussion in the form of a keynote, paper presentations, and an expert panel discussion (morning session). This will help connect participants that share similar interests and provide them with a chance to contribute and learn.

The morning session will begin with a keynote by leading expert in human-AI collaboration. Participants will have the opportunity to share their accepted work with either paper or poster presentations. The morning session will include two paper sessions for spotlight authors to share their accepted work. The presentations will consist of 5-7 minutes lightning talks, followed by a joint (around 10-minute) Q&A session. These talks will be pre-recorded; Non-spotlight acceptances will also have pre-recorded videos at 1-3 minute lengths. We will post all these recordings online for asynchronous access since 2022 participants found it overwhelming to play all videos during paper sessions. Concurrent with the planned coffee breaks (in the morning and afternoon sessions), we also plan to hold poster presentations to support more interactions between authors and participants and play videos of the 1-3 minute presentations.

We will also host a discussion panel of experts with balanced perspectives from academia and industry, to form the discussion around our diverse research interests—trust calibration, human-AI teaming, understanding AI uncertainty, the evolving role of humans in human-AI collaboration, etc. We have commitments from five experts from academia and industry with expertise in human-AI collaboration: Stephanie Bell (Research Scientist at the Partnership on AI with expertise on future of work and the evolving human role in human-AI partnerships), Michael Bernstein (Professor of Computer Science at Stanford with expertise in human-AI interaction), Su Lin Blodgett (Senior Researcher at Microsoft with expertise on language technologies), Jina Suh (Principal Researcher at Microsoft Research with expertise in workplace wellbeing).

In the afternoon, we plan to allocate adequate time for two sessions of in-depth group activities, each session containing a one-hour within-group discussion, and a half-hour between-grouping insight sharing. The groups will be in the form of "birds-of-feather" discussions and practices around several topics, including measures, challenges, and mechanisms and interactions for shaping trust. We will finalize the group activities based on the number of participants and their interests, but some initial ideas include,

(1) a *debating format*, where two groups are paired to represent the claims and counterclaims relating to themes within human-AI trust, so to motivate people to play devil's advocates to each others ideas. The organizers would provide inspirational questions, as well as imaginary use scenarios that can ground these discussions (e.g., in high-stake domains like education, medical, etc.)
(2) *concept mapping* around definitions, measures, and factors for appropriate trust and reliance. Groups could collaborate to enumerate and discuss relevant concepts—predictability, uncertainty, trust, reliance, adaptability, etc. and identify their overlaps and relationships.
(3) *ideation* for solutions for shaping trust for particular use cases, such as news recommender systems or autonomous vehicles; here groups can brainstorm possible solutions, including system interactions, explanations, or visualizations, and iterate on these ideas with input from other groups. Outputs of this activity might consist of a set of solution ideas or low-fidelity mockups for system designs.
(4) *on-the-spot paper writing and reviewing*: where participants come up with one research idea, or an imaginary paper they would like to write around the topic of trust and reliance. Groups' output might be an abstract, certain teaser figures illustrating the core idea, or compelling use cases. Then, participants will review these deliveries, hopefully to help better articulate what aspects people would care about around a particular research idea.

Each group will be moderated by at least one organizer; we will also encourage paper authors to join groups related to their paper topics and share their posters within the group, so they can have more dedicated discussions around the broader theme, but in the context of their own work. For hybrid participation, activities will make use of collaborative virtual environments like Google Documents and Miro boards.

We will host the paper lightning talks, posters, and group sharings in Google Slides and on the website, and we will later convert them into medium posts to share with the broader audience.

6 POST-WORKSHOP PLAN

We will synthesize our findings from the workshop as a Medium post. In order to reach a larger audience, we will upload the recorded sessions on YouTube, publish group activity outcomes as Medium posts, and share all the materials on social media.

7 CALL FOR PARTICIPATION: WORKSHOP ON TRUST AND RELIANCE IN EVOLVING AI-HUMAN WORKFLOWS

State-of-the-art AIs and LLMs (e.g., GPT-4) can now perform tasks previously exclusive to humans (e.g., writing code, generating ideas, planning), and are being widely used in various domains, applications, and commercial tools (e.g., GitHub Copilot, Bing Chat, Bard, ChatGPT, Advanced Data Analytics, etc). Compared to more traditional forms of specialized AI, these LLMs introduce new dynamics to human-AI interaction. For example, instead of simply trusting and depending on AI models as supportive tools that offer single-shot suggestions, humans may increasingly come to rely on AIs as *collaborative peers* and engage in dialogs and negotiations, delegate various tasks. Given these impressive capabilities and evolving landscape of LLMs, what will the new human roles in human-AI collaboration become? how will we adopt and rely on these models differently? As these systems move to large-scale adoption, how will people's skills on the task be impacted? Will AI assistance help or hinder people's skill improvement over time?

This workshop will provide a venue for exploring how the collaboration between humans and AIs evolve, and how humans' trust and reliance on AIs evolve accordingly. We invite participants with expertise in HCI, AI, ML, psychology, and social science, or other relevant fields to come together and discuss three themes: (1) What should be the new roles of humans and AIs when the latter are so versatile? (2) How should we define and measure trust and reliance in these new contexts? (3) What are the long term impact of such human-AI collaboration on each other?

Themes include, but are not limited to:

- Investigating how humans adapt to being collaborators and decision-makers alongside AI systems.
- Assessing the challenges and opportunities for individuals in evolving AI-driven work environments.
- Investigating the psychological aspects of trust when working with LLMs and AI systems as collaborative peers.
- Analyzing the impact of factors like transparency, explainability, and system performance on trust and reliance.
- Examining the sustained effects of using AI assistance on human skills and expertise.

The submission should use ACM single column format, and should take 4-10 pages. At least one author must register and attend the workshop. Submission will be reviewed by program committee and accepted papers will be posted on the workshop website and shared via social media.

Important Dates:

- Submission: February 23, 2024 (Easychair)
- Notifications: March 14, 2024
- Camera Ready: April 11, 2024
- Workshop: May 11, 2024

REFERENCES

[1] Gagan Bansal, Besmira Nushi, Ece Kamar, Eric Horvitz, and Daniel S Weld. 2021. Is the most accurate ai the best teammate? optimizing ai for teamwork. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 11405–11414.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712 [cs.CL]

[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] Erin K Chiou and John D Lee. 2021. Trusting Automation: Designing for Responsivity and Resilience. *Human Factors* (2021), 00187208211009995.

[5] Tyna Eloundou, Sam Manning, Pamela Mishkin, and Daniel Rock. 2023. Gpts are gpts: An early look at the labor market impact potential of large language models. *arXiv preprint arXiv:2303.10130* (2023).

[6] 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.

[7] Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, et al. 2022. Predictability and surprise in large generative models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 1747–1764.

[8] Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd-workers for text-annotation tasks. *arXiv preprint arXiv:2303.15056* (2023).

[9] Ella Glikson and Anita Williams Woolley. 2020. Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals* 14, 2 (2020), 627–660.

[10] Patrick Hemmer, Sebastian Schellhammer, Michael Vössing, Johannes Jakubik, and Gerhard Satzger. 2022. Forming effective human-AI teams: building machine learning models that complement the capabilities of multiple experts. *arXiv preprint arXiv:2206.07948* (2022).

[11] Patrick Hemmer, Monika Westphal, Max Schemmer, Sebastian Vetter, Michael Vössing, and Gerhard Satzger. 2023. Human-AI Collaboration: The Effect of AI Delegation on Human Task Performance and Task Satisfaction. In *Proceedings of the 28th International Conference on Intelligent User Interfaces*. 453–463.

[12] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. 2021. Unsolved problems in ml safety. *arXiv preprint arXiv:2109.13916* (2021).

[13] Lea Krause and Piek Vossen. 2020. When to explain: Identifying explanation triggers in human-agent interaction. In *2nd Workshop on Interactive Natural Language Technology for Explainable Artificial Intelligence*. 55–60.

[14] John D Lee and Katrina A See. 2004. Trust in automation: Designing for appropriate reliance. *Human factors* 46, 1 (2004), 50–80.

[15] Ariel Levy, Monica Agrawal, Arvind Satyanarayan, and David Sontag. 2021. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[16] Zhuoran Lu and Ming Yin. 2021. Human Reliance on Machine Learning Models When Performance Feedback is Limited: Heuristics and Risks. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–16.

[17] Qianou Ma, Tongshuang Wu, and Kenneth R Koedinger. 2023. Is AI the Better Programming Partner. *Human-Human Pair Programming vs. Human-AI pAIr Programming. CoRR, abs/2306.05153* (2023).

[18] Hussein Mozannar, Gagan Bansal, Adam Fourney, and Eric Horvitz. 2022. Reading between the lines: Modeling user behavior and costs in AI-assisted programming. *arXiv preprint arXiv:2210.14306* (2022).

[19] Namkee Oh, Gyu-Seong Choi, and Woo Yong Lee. 2023. ChatGPT goes to the operating room: evaluating GPT-4 performance and its potential in surgical education and training in the era of large language models. *Annals of Surgical Treatment and Research* 104, 5 (2023), 269.

[20] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[21] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. arXiv:2203.02155 [cs.CL]

[22] Vanessa Sauer, Alexander Mertens, Jens Heitland, and Verena Nitsch. 2021. Designing for Trust and Well-being: Identifying Design Features of Highly Automated Vehicles. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–7.

[23] Leslie Y Garfield Tenzer. 2023. Defamation in the Age of Artificial Intelligence. *Available at SSRN 4545070* (2023).

[24] Michel Wermelinger. 2023. Using GitHub Copilot to solve simple programming problems. In *Proceedings of the 54th ACM Technical Symposium on Computer Science Education V. 1*. 172–178.

[25] David Gray Widder, Laura Dabbish, James D Herbsleb, Alexandra Holloway, and Scott Davidoff. 2021. Trust in Collaborative Automation in High Stakes Software Engineering Work: A Case Study at NASA. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–13.

[26] Kyle Wiggers. 2022. Copilot, GitHubs AI-powered programming assistant, is now generally available. *TechCrunch* (2022).