Proceedings of the Eighth AAAI/ACM Conference on AI, Ethics, and Society (AIES 2025)

Human Centered AI for Research Ethics and Transparency

Tatiana Chakravorti[1],

[1]Pennsylvania State University, USA
tfc5416@psu.edu

Abstract

In recent years, empirical sciences have faced a crisis of confidence, sparked by high-profile replication projects that failed to reproduce key findings. This "replication crisis" has prompted widespread reflection and a growing movement advocating for reform. Central to this movement is the belief that reproducibility, transparency, and rigorous evaluation are essential to scientific progress. Our research aligns with these goals, emphasizing how emerging technologies can offer innovative tools to address challenges in scientific integrity and reliability. This project outlines a research agenda focused on the use of artificial intelligence (AI) and hybrid human-AI systems to assess the replicability of scientific findings and to better understand science as a socially embedded practice. We develop and test a new class of AI-driven prediction markets, where algorithmic agents trade contracts tied to the outcomes of replication studies. Recognizing the value of human judgment, we explore hybrid prediction markets that integrate human participants alongside AI agents. These systems are examined through simulations and pilot experiments in real-world settings. To guide the design of these tools, we conduct in-depth surveys and interviews with researchers, gathering insights into the practical needs and concerns surrounding AI and hybrid systems in scientific workflows. Particular attention is paid to transparency, explainability, and the social and cultural dimensions that shape scientific practices. Our work aims to inform the development of technologies that not only support more reliable science but also respect and reflect the values of the scientific community.

1 Introduction

Motivation Reproducibility, replicability, and transparency are cornerstones of scientific integrity, ensuring that research findings are trustworthy and robust. Open science practices such as sharing data, code, and methodologies have become central to advancing these principles, enabling other researchers to validate and build upon existing work. In the past decade, the open science and science of science communities have made significant strides in addressing concerns about the credibility of published findings. However, substantial gaps remain. Many researchers are still unaware of the challenges posed by the reproducibility crisis or are uncertain about best practices for producing reliable, transparent research. In the context of rapidly evolving artificial intelligence (AI) technologies, this study explores how AI can be harnessed to support and improve scientific credibility. At the same time, it recognizes the importance of explainability in AI systems, particularly when applied to high-stakes tasks such as evaluating scientific validity. The core aim of this work is to foster a more open, interpretable, and globally accessible scientific process—where research outputs are not only available but also understandable and usable across disciplines.

To address the replication crisis, we investigate human-AI collaborative approaches that integrate computational power with human judgment. While AI systems have shown great promise, they continue to face limitations in areas requiring nuanced reasoning, creativity, and learning from limited data. These challenges highlight the need for hybrid strategies that combine the strengths of both humans and machines. By developing and evaluating such collaborative frameworks, this study seeks to improve our ability to anticipate the replicability of scientific claims and enhance trust in the research enterprise.

The goal of my research work is to address three major issues in scientific integrity:

1. Exploring Researchers' Perceptions of Integrity and AI in Science This objective seeks to understand how researchers across disciplines perceive the current state of research integrity, as well as their attitudes toward the use of AI-driven replicability prediction tools. Through surveys and interviews, we investigate not only the perceived usefulness and trustworthiness of such technologies, but also the social, cultural, and epistemic factors shaping their acceptance and implementation in the scientific workflow.

2. Advancing Scientific Integrity through Hybrid Prediction Markets This objective focuses on designing and evaluating hybrid prediction markets that combine human expertise with AI agents to forecast the replicability of published scientific findings. By leveraging the complementary strengths of human intuition and machine learning, the study aims to create scalable, reliable tools that can support meta-scientific evaluation and foster a culture of accountability and evidence-based confidence in research outputs.

3. Investigating Ethical AI Use in Peer Review Ecosystems This objective examines the ethical implications of

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

deploying large language models (LLMs), such as ChatGPT, in research and peer review processes. By engaging with reviewers, journal editors, and conference organizers, we explore how LLMs might augment or undermine traditional scholarly roles, and how principles of Open Science—such as transparency, accessibility, and collaborative scrutiny—can guide the responsible use of AI. This includes identifying risks (e.g., overreliance, opacity, bias), and opportunities (e.g., improved reviewer efficiency, enhanced editorial consistency, democratized access to expertise).

Human Centered AI for Research Ethics The primary goal of this work is to contribute to ongoing efforts to address the scientific credibility crisis by developing and validating novel AI methods for assessing scientific findings, promoting open science practices, and enhancing collaboration between humans and AI in the research process.

While AI and machine learning (ML) have made remarkable progress in achieving human-level or even superhuman performance in a range of complex tasks, their practical deployment continues to face critical challenges. AI models often lack common sense and struggle with tasks requiring creativity, intuition, or learning from limited data. Moreover, although established classification algorithms have demonstrated success, they have not yet fully earned user trust in high-stakes, decision-making contexts. Drawing on a *mixed-methods* approach that includes surveys followed by semi-structured interviews, we explore how social science researchers navigate excitement and tensions around automation when working with AI and ML. Participants raised concerns about de-skilling, lack of standardization, ethical risks, bias, model interpretability, inadequate training, and over-reliance on AI systems. *Ethical concerns were less pronounced when we asked about ML techniques than when we asked about AI, suggesting evolution in the distribution of attention with the rise of GenAI.*

These limitations have led to growing interest in hybrid human-AI approaches that seek to integrate the strengths of both human reasoning and artificial intelligence. In this study, we conducted extensive experiments with an artificial prediction market model to investigate how various market parameters influence model performance on benchmark classification tasks. We further demonstrated, through simulation, the effects of introducing exogenous agents designed to represent primitive human behaviors on market dynamics and predictive outcomes.(Chakravorti et al. 2023b).

A prototype human-AI prediction market is built, showcasing its potential for facilitating meaningful human-AI collaboration(Chakravorti et al. 2023a). Within these markets, AI agents (bot traders) are trained to buy outcomes of future events alongside humans. Such classification decisions can be interpreted as outcomes of these events, with the price of an asset linked to a specific classification outcome serving as an indicator of the system's confidence in that decision. By integrating human participants into these markets alongside bot traders, we amalgamate insights from both sources. This research work was conducted with prototype hybrid markets aimed at predicting the outcomes of replication studies. We explored both the challenges and opportunities presented by this approach, shared insights from semi-structured interviews with participants of the hybrid market, and provided a vision for future research and development in this area.

In this work, we focused on the formative evaluation of the replication prediction tool, which is the product of my artificial prediction market(Chakravorti et al. 2023c; Wu et al. 2024). Formative evaluation is carried out during the development or improvement phase of the subject under evaluation, with the primary goal of gathering feedback for real-time enhancements. We conducted semi-structured interviews with faculty members and PhD scholars from universities across India to gauge their current research practices and their understanding of the reproducibility crisis. There are two groups of researchers, one is social science and the one is engineering, to explore both perspectives. The prototype AI replication prediction tool will be introduced as a basis for discussions on how computational tools can be designed and implemented to support their work with research ethics. By centering these perspectives, we explore the acceptance and perceptions of hybrid human-AI solutions and also how they can integrate this tool into their research.

Finally, we have made lots of efforts to understand experiences and perspectives surrounding reproducibility, replicability, and open science, and have predominantly focused on researchers and research communities in Western countries. This research work represents an initial step towards incorporating cultural perspectives by conducting a comparative study of researchers in the USA and India(Chakravorti, Koneru, and Rajtmajer 2025). The goal is to conduct a comparative analysis of the state of reproducibility and open science in India and the USA, identifying the challenges researchers face in these countries, the incentives necessary for fostering transparent research practices, the features researchers look for in a paper to assess its credibility, and their perceptions and experiences with the current peer review process.

2 Conclusion

This research provides significant contributions to the fields of science of science, human-AI collaboration, and ethical research practices. It offers innovative solutions to long-standing challenges in reproducibility and peer review, advocates for the integration of cultural perspectives, and explores the frontier of human-AI collaboration in advancing scientific integrity and transparency.

References

Chakravorti, T.; Fraleigh, R.; Fritton, T.; McLaughlin, M.; Singh, V.; Griffin, C.; Kwasnica, A.; Pennock, D.; Giles, C. L.; and Rajtmajer, S. 2023a. A prototype hybrid prediction market for estimating replicability of published work. In *HHAI 2023: Augmenting Human Intellect*, 300–309. IOS Press.

Chakravorti, T.; Koneru, S.; and Rajtmajer, S. 2025. Reproducibility and replicability in research: What 452 professors think in Universities across the USA and India. *PloS one*, 20(3): e0319334.

Chakravorti, T.; Singh, V.; Rajtmajer, S.; McLaughlin, M.; Fraleigh, R.; Griffin, C.; Kwasnica, A.; Pennock, D.; and Giles, C. L. 2023b. Artificial Prediction Markets Present a Novel Opportunity for Human-AI Collaboration. In *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems*, 2304–2306.

Chakravorti, T.; Wu, C.; Koneru, S.; and Rajtmajer, S. 2023c. Perspectives from India: Opportunities and Challenges for AI Replication Prediction to Improve Confidence in Published Research. *arXiv preprint arXiv:2310.19158*.

Wu, C.; Chakravorti, T.; Carroll, J. M.; and Rajtmajer, S. 2024. Integrating measures of replicability into scholarly search: Challenges and opportunities. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, 1–18.