Fairness Agents in Scientific Collaboration: A Research Agenda

Author 1 **Author 2**

Abstract

This paper introduces the concept of Fairness Agents: autonomous software agents embedded in scientific workflows to detect, explain, and mitigate bias in collaborative knowledge production. While algorithmic fairness has primarily focused on predictive models, scientific collaboration involves complex interpersonal and institutional processes where bias often arises. We identify a research gap at the intersection of multi-agent systems, epistemic justice, and AI fairness. Drawing on a structured literature synthesis, we define Fairness Agents, propose a typology (observer, interventionist, and reflective agents), and outline a functional architecture. We illustrate their relevance through use cases in interdisciplinary research, peer review, and open science. The paper concludes by discussing key design challenges—transparency, trust, and norm conflict—and proposes directions for future evaluation and participatory co-design. Fairness Agents offer a path toward more inclusive and accountable agent-mediated science.

1 Introduction & Theoretical Background

Scientific discovery increasingly relies on AImediated workflows: from hypothesis generation to data analysis, peer review, and publication. The new Agents4Science conference encourages work exploring how AI agents can autonomously author and review scientific contributions, offering a radically transparent experimental sandbox for AI-driven science. Within this emergent landscape, a crucial but underexplored question arises: How can agents help uphold fairness in collaborative scientific processes?

Bias in science is multifold: underrepresentation of marginalized groups in research participation, uneven credit attribution, epistemic exclusion, and datadependent inequities. Scientific collaborations—especially in interdisciplinary or health domains—can reinforce these dynamics if unchecked. Although fairness in AI models has received increasing attention—from formal definitions like demographic parity, sufficiency, or counterfactual fairness to humanintheloop mitigation frameworks—scientific workflows lack embedded mechanisms for fairness auditing.

Meanwhile, growing interest in multiagent fairness auditing shows promising results: coordinated agents auditing a shared platform achieve more accurate detection than isolated audits, although excessive coordination can be counterproductive. Yet such frameworks focus on model fairness in application contexts—not on fairness among collaborating agents or between agents and humans within scientific teams.

Existing research on algorithmic fairness tends to focus on modelcentric outcomes (e.g., balanced error rates, causal mediation), missing the broader dynamics of how agents interact with each other and with human collaborators in a scientific setting. There is currently no formal concept of Fairness Agents: autonomous entities designed to observe, detect, explain, and intervene on fairness issues across tasks, credit assignment, data provenance, and inclusion within agent-mediated scientific ecosystems.

We therefore propose to introduce and formalize the notion of Fairness Agents—autonomous agents whose purpose is to monitor procedural and representational fairness in scientific collaborations,

Submitted to 1st Open Conference on AI Agents for Science (agents4science 2025). Do not distribute.

Type	Role	Example Functions	Intervention Mode
Observer Agent	Passive monitor	Track speaking time, data provenance, author contribution patterns	Signal alerts or visualize inequalities
Interventionist Agent	Active corrector	Recommend inclusion of missing perspectives; block biased workflows; rebalance authorship	Interrupt or redirect agent/human decisions
Reflective Agent	Contextual explainer	Generate fairness reports; trace bias origins; assess epistemic diversity	Foster group reflection and documentation

Table 1: Typology of Fairness Agents in scientific collaboration.

particularly under conditions of interdisciplinary work, intersectional bias, and epistemic asymmetry. Our core thesis is:

Fairness Agents can operate throughout agent-mediated scientific workflows—acting as auditors, explainers, and corrective nudgers—to systematically detect and mitigate fairness violations while preserving epistemic productivity.

Research Question: What roles can fairness-oriented agents play in identifying and mitigating bias in collaborative scientific workflows, and what design challenges must be addressed to integrate them effectively?

2 Methodology

This study adopts a conceptual research design grounded in theory synthesis and design-oriented reasoning. Our aim is not to evaluate a technical implementation, but rather to introduce and refine a new conceptual construct—the Fairness Agent—and to articulate its potential roles, functions, and challenges within agent-mediated scientific collaboration.

The core method involves drawing on existing, interdisciplinary literature to systematically construct a coherent conceptual model. In doing so, we identify patterns, gaps, and tensions across research on AI fairness, multi-agent systems, and the sociology of science.

Literature strategy:

A seed corpus of ~40 key publications was assembled from AI fairness, science and technology studies (STS), and agent-based systems.

This was expanded via snowballing and database searches using terms like "multi-agent fairness," "epistemic exclusion in science," and "AI in peer review."

Sources included conference proceedings (FAccT, CHI, AAMAS), journal articles, and open preprints.

Analytical steps:

Identify fairness-relevant agent functions and system roles

Map failure modes in scientific collaboration (e.g., epistemic bias, credit asymmetry)

Develop a role typology and functional architecture

Propose use cases and agenda for evaluation

Limitations: As a conceptual paper, this work is not empirically validated but forms the groundwork for future implementation, simulation, and user research.

3 Results

4 2 Typology of Fairness Agents

5 3 Functional Architecture

Interaction Layer: logs agent-human interactions and communication patterns

Data Layer: accesses metadata, provenance, and datasets for auditing

Governance Layer: embeds soft/hard rules for fairness enforcement or nudging

6 4 Use Cases

Health research teams: detect exclusion of minoritized disciplines in interdisciplinary projects

Peer review platforms: assess reviewer bias and uneven evaluation standards

Open science consortia: ensure credit and resource access equity across institutions

7 Discussion

This paper extends the fairness discourse from predictive models to the social and epistemic infrastructures of science. While fairness auditing tools exist for outputs, they are inadequate for managing fairness as a process within collaborative ecosystems. Fairness Agents offer a mechanism for embedding fairness principles directly into scientific workflows.

Our typology emphasizes multiple levels of engagement—from passive tracking to active policy enforcement to reflective reporting—aligning with literature on epistemic justice, value-sensitive design, and participatory AI governance. These agents can support trust, procedural accountability, and inclusion—but only if carefully aligned with human values and domain norms.

Key design challenges include:

Trust and transparency: interventions must be explainable and auditable

Norm conflict resolution: agents will encounter competing fairness norms (e.g., equity vs. meritocracy)

Avoiding bias-by-design: agent goals must be inclusive and reflexively designed

8 Conclusion

We introduced the concept of Fairness Agents—autonomous agents embedded in scientific workflows to support epistemic inclusion and procedural fairness. Our contributions include:

A typology of agent roles

A functional architecture

Practical use case scenarios

Future work should focus on:

Agent-based simulations

Participatory design with diverse research communities

Evaluative criteria for epistemic and procedural fairness in scientific AI systems

Fairness agents represent a critical step toward more inclusive, reflexive, and socially responsible science in the age of autonomous agents.

9 References

Ananthanarayanan, R., et al. (2023). Autonomous discovery in materials science with large language models. Nature Reviews Materials, 8(2), 123–135.

Bardzell, S. (2010). Feminist HCI: Taking stock and outlining an agenda for design. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, 1301–1310.

Barocas, S., Hardt, M., & Narayanan, A. (2019). Fairness and Machine Learning: Limitations and Opportunities. fairmlbook.org.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. Proceedings of the 2018 Conference on Fairness, Accountability, and Transparency, 149–159.

Chen, T., et al. (2023). Towards autonomous scientific research agents. arXiv preprint arXiv:2306.03666.

D'Ignazio, C., & Klein, L. F. (2020). Data Feminism. MIT Press.

Dotson, K. (2014). Conceptualizing epistemic oppression. Social Epistemology, 28(2), 115–138.

Fricker, M. (2007). Epistemic Injustice: Power and the Ethics of Knowing. Oxford University Press.

Friedman, B., Kahn, P. H., & Borning, A. (2006). Value sensitive design and information systems. Human-computer interaction and management information systems: Foundations, 1, 348–372.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. Advances in Neural Information Processing Systems, 29, 3315–3323.

Hevner, A. R., et al. (2004). Design science in information systems research. MIS Quarterly, 28(1), 75–105.

Holstein, K., et al. (2019). Improving fairness in machine learning systems: What do industry practitioners need? CHI Conference on Human Factors in Computing Systems, 1–16.

Jacovi, A., et al. (2021). Formalizing trust in artificial intelligence. CHI Conference, 1–18.

Kearns, M., et al. (2019). An empirical study of rich subgroup fairness. FAT, 100–109.

Klar, T., et al. (2024). Multi-agent fairness auditing: A simulation study. arXiv preprint arXiv:2402.08522.

Lee, M. K., et al. (2022). Working with machines: Impact of algorithmic management. ACM Transactions on Computer-Human Interaction, 29(1), 1–34.

Medina, J. (2013). The Epistemology of Resistance. Oxford University Press.

Mitchell, M., et al. (2019). Model cards for model reporting. FAT, 220–229.

Park, J., et al. (2023). Generative agents: Interactive simulacra of human behavior. arXiv preprint arXiv:2304.03442.

Rahwan, I., et al. (2019). Machine behaviour. Nature, 568(7753), 477–486.

Raji, I. D., et al. (2020). Closing the AI accountability gap. FAT, 33–44.

Selbst, A. D., et al. (2019). Fairness and abstraction in sociotechnical systems. FAT, 59–68.

Sloane, M., et al. (2022). Participation is not a design fix. FAT, 677–689.

Wooldridge, M. (2009). An Introduction to MultiAgent Systems. Wiley.

Zou, J. Y., et al. (2023). Science as a multi-agent system. Patterns, 4(5), 100723.

Agents4Science AI Involvement Checklist

1. **Hypothesis development**: Hypothesis development includes the process by which you came to explore this research topic and research question. This can involve the background research performed by either researchers or by AI. This can also involve whether the idea was proposed by researchers or by AI.

   Answer: **[C]**

   Explanation: The thematic area was proposed by the human author, but the LLM generated detailed subquestions and conceptual directions that were reviewed and structured by the human author.

2. **Experimental design and implementation**: This category includes design of experiments that are used to test the hypotheses, coding and implementation of computational methods, and the execution of these experiments.

   Answer: **[C]**

   Explanation: As a conceptual paper, design referred to structuring the framework and selecting literature. The LLM proposed candidate structures and sequences; the human author guided, constrained, and approved them.

3. **Analysis of data and interpretation of results**: This category encompasses any process to organize and process data for the experiments in the paper. It also includes interpretations of the results of the study.

   Answer: **[D]**

   Explanation: No empirical data were analysed; literature synthesis and conceptual interpretation were drafted entirely by the LLM and then critically reviewed and accepted by the human author.

4. **Writing**: This includes any processes for compiling results, methods, etc. into the final paper form. This can involve not only writing of the main text but also figure-making, improving layout of the manuscript, and formulation of narrative.

   Answer: **[C]**

   Explanation: The writing was mainly done by AI. Human assistance was used to add the subchapter titles, as this division was too difficult for the AI.

5. **Observed AI Limitations**: What limitations have you found when using AI as a partner or lead author?

   Description: We observed occasional hallucinated or imprecise citations, shallow synthesis of complex literatures, and conflation of adjacent conceptual domains. These issues required human curation, clarification of constructs, and iterative editing to preserve conceptual precision.

Agents4Science Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: answerYes

   Justification: The abstract and introduction accurately reflect the paper's conceptual contributions (definition, typology, architecture, use cases).

   Guidelines:

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: answerYes

   Justification: We explicitly note LLM-specific issues (hallucinated references, shallow synthesis, domain conflation) and conceptual scope limits.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: answerNA

   Justification: No empirical experiments were conducted; this is a conceptual research agenda.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: answerNA

   Justification: No dataset or code were produced in this conceptual work.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

   Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the Agents4Science Code of Ethics (see conference website)?

   Answer: [NA]

   Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Not applicable for a conceptual paper; no empirical experiments, datasets, or models are introduced.