Geoff Keeling

London, United Kingdom

/in/geoff-keeling
geoffkeeling.github.io
geoffkeelingbhyat@gmail.com

Employment

Staff Research Scientist, *Google Research*. 2025–Present
Senior Research Scientist, *Google Research*. 2023–25
Research Scientist, *Google Research*. 2023–23
Bioethicist, *Google Health*. 2021–23
Postdoctoral Fellow, Institute for Human-Centered AI & Center for Ethics in Society, *Stanford University*. 2020–21
Research Assistant, Leverhulme Centre for the Future of Intelligence, *University of Cambridge*. 2019–20

Affiliations

Fellow, Institute of Philosophy, School of Advanced Study, *University of London*. 2024–Present
Associate Fellow, Leverhulme Centre for the Future of Intelligence, *University of Cambridge*. 2020–Present

Education

PhD Philosophy, *University of Bristol*, No Corrections. 2017–20
MA Philosophy, *University of Bristol*, Distinction in all Components (Rank: 1st/24). 2016–17
BSc Philosophy, Logic, and Scientific Method, *London School of Economics*, First Class (Rank: Joint 1st/51). 2013–16

Research

Ethics of AI/ML, Moral and Political Philosophy, Bioethics, Cognitive Science.

Papers

1. Gabriel, I. & **Keeling, G.** (2025). A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, 1-23.
2. **Keeling, G.** & Street, W. (2025). On the attribution of confidence to large language models. *Inquiry*, 1-27.
3. Broestl, N.*, Lange, B.*, Voinea, C.*, **Keeling, G.***, & Lam, R. (2025). Evaluating intra-firm LLM alignment strategies in business contexts. *arXiv preprint arXiv:2505.18779v1.* (*Equal Contribution)
4. **Keeling, G.***, Street, W.*, Stachaczyk, M., Zakharova, D., Comșa, I., Sakovych, A., Logothetis, I., Zhang, Z., Agüera y Arcas, B., & Birch, J. (2024). Can LLMs make trade-offs involving stipulated pain and pleasure states? *arXiv preprint arXiv:2411.02432v1.* (*Equal Contribution) [Media: Scientific American, Futurism, The Guardian, MSN]
5. Gabriel, I.*, Manzini, A.*, **Keeling, G.***, ... , Agüera y Arcas, B., Isaac, W., & Manyika, J. (2024). The ethics of advanced AI assistants. *arXiv preprint arXiv:2404.16244.* (*Equal Contribution) [Media: WIRED, CBS, The Information, Axios, The Verge, MSN, Tech Policy Press, MarkTechPost]

6. Manzini, A., **Keeling, G.**, Marchal, N., McKee, K., Reiser, V., Haas, J., & Gabriel, I. (2024, June). Should users trust advanced AI assistants? Justified trust as a function of competence and alignment. In: *Proceedings of FAccT '24: ACM Conference on Fairness, Accountability, and Transparency.*

7. Manzini, A., **Keeling, G.**, Alberts, L., Vallor, S., Ringel Morris, M., & Gabriel, I. (2024, October). The code that binds us: Navigating the appropriateness of human-AI assistant relationships. In: *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society.* [🏆 Runner Up: Best Paper Award]

8. El-Sayed, S., Akbulut, C., McCroskery, A., **Keeling, G.**, Kenton, Z., Jalan, Z., Marchal, N., Manzini, A., Shevlane, T., Vallor, S., Susser, D., Franklin, M., Bridgers, S., Law, H., Rahtz, M., Shanahan, M., Tessler, M.H., Douillard, A., Everitt, T., & Brown, S. (2024). A mechanism-based approach to mitigating harms from persuasive generative AI. *arXiv preprint arXiv:2404.15058.* [Media: VentureBeat, Futurism]

9. Street, W., Siy, J.O., **Keeling, G.**, Baranes, A., Barnett, B., McKibben, M., Kanyere, T., Lentz, A., Agüera y Arcas, B., & Dunbar, R. (2024). LLMs achieve adult human performance on higher-order theory of mind tasks. *arXiv preprint arXiv:2405.18870v1.*

10. Schaekermann, M., Spitz, T., Pyles, M., Cole-Lewis, H., Wulczyn, E., Pfhol, S.R., Martin, D., Jaroensri, R., **Keeling, G.**, Liu, Y., Farquhar, S., Xue, Q., Lester, J., Hughes, C., Strachan, P., Tan, F., Bui, P., Mermel, C.H., Peng, L.H., Matias, Y., Corrado, G.S., Webster, D.R., Virmani, S., Semturs, C., Liu, Y., Horn, I., & Chen, P-H.C. (2024). Health equity assessment of machine learning performance (HEAL): a framework and dermatology AI model case study. *Lancet eClinicalMedicine.* [Media: MarkTechPost]

11. Alberts, L., **Keeling, G.**, & McCroskery, A. (2024). Should agentic conversational AI change how we think about ethics? Characterising an interactional ethics centred on respect. *arXiv preprint arXiv:2401.09082v2.*

12. **Keeling, G.**, Lange, B., McCroskery, A., Pedersen, K., Weinberger, D., & Zevenbergen, B. (2024). Moral imagination for engineering teams: The technomoral scenario. *International Review of Information Ethics*, 34(1).

13. Lange, B.*, **Keeling, G.***, McCroskery, A., Zevenbergen, B., Blascovich, S., Pedersen, K., Lentz, A., & Agüera y Arcas, B. (2023). Engaging engineering teams through moral imagination: a bottom-up approach for responsible innovation and ethical culture change in technology companies. *AI and Ethics*, 1-10. (*Equal Contribution)

14. **Keeling, G.** (2023). Algorithmic bias, generalist models, and clinical medicine. *AI and Ethics*, 1-12.

15. **Keeling, G.** & Nyrup R. Explainable machine learning, patient autonomy, and clinical reasoning. In: Véliz, C. (Ed). (2023). *Oxford Handbook of Digital Ethics* (pp. 528–552). Oxford University Press.

16. **Keeling, G.** (2023). A dilemma for reasons additivity. *Economics & Philosophy*, 39(1), 20-42.

17. Grote, T. & **Keeling, G.** (2022). Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, *24(3)*, 39.

18. **Keeling, G.** Automated vehicles and the ethics of classification. In: Jenkins, R., Cerný, D., & Hríbek, T. (Eds). (2022). *Autonomous Vehicle Ethics: The Trolley Problem and Beyond* (pp. 41-57). Oxford University Press.

19. Grote, T. & **Keeling, G.** (2022). On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics*, *31*(1), 83-94.

20. **Keeling, G.** & Burr, C. Digital manipulation and mental integrity. In: Jongepier, F., & Klenk, M. (2022). *The Philosophy of Online Manipulation* (pp. 253-271). Routledge.

21. **Keeling, G.** & Paterson, N. (2022). Proper functions: etiology without typehood. *Biology & Philosophy*, 37(3), 19.

22. Bommasani, R., ... , **Keeling, G.**, ... , & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258.* [Media: Forbes, The Economist, VentureBeat, WIRED]

23. **Keeling, G.** (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, *26*(1), 293-307.

24. **Keeling, G.**, Evans, K., Thornton, S. M., Mecacci, G., & Santoni de Sio, F. Four perspectives on what matters for the ethics of automated vehicles. In: Meyer, G. & Beiker, S. (Eds). (2019). *Road Vehicle Automation 6* (pp. 49-60). Springer International Publishing.

25. **Keeling, G.** (2018). Legal necessity, Pareto efficiency & justified killing in autonomous vehicle collisions. *Ethical Theory and Moral Practice*, *21*(2), 413-427.

26. **Keeling, G.** Against Leben's Rawlsian collision algorithm for autonomous vehicles. In: Müller, V. (Ed). (2018). *Philosophy and Theory of Artificial Intelligence 2017* (pp. 259-272). Springer International Publishing.

27. **Keeling, G.** (2018). The sensitivity argument against child euthanasia. *Journal of Medical Ethics*, *44*(2), 143-144.

28. **Keeling, G.** (2018). Autonomy, nudging and post-truth politics. *Journal of Medical Ethics*, 44(10), 721-722.

29. **Keeling, G.** (2017). Commentary: Using virtual reality to assess ethical decisions in road traffic scenarios: applicability of value-of-life-based models and influences of time pressure. *Frontiers in Behavioral Neuroscience*, 11, 247-48.

Talks

1. **Panel: AI and Moral Patienthood**, *with Henry Shevlin and Winnie Street*.
   AI, Animals and Digital Minds Conference, London, 2025.

2. **Animal-inspired Approaches to Assessing LLM Sentience**, *with Winnie Street*.
   AI Welfare Workshop, Eleos AI, Berkeley, 2025.

3. **Ethics and the Agentic Turn**.
   AI Ethics, Risks and Safety Conference, Bristol, 2025.

4. **What About the Moral Claims of AI Agents?**.
   Symposium: Themes from Iason Gabriel, Department of Philosophy, Hong Kong University, 2025.

5. **Do Large Language Models Have Credences?**
   AI and Ethics Workshop, Princeton University, 2024.
   Institute for Ethics in Technology, Hamburg University of Technology, 2024.

6. **Can LLMs Make Trade-Offs Involving Stipulated Pain and Pleasure States?** *with Winnie Street.*
   Institute of Philosophy, School of Advanced Study, University of London, 2024.

7. **The Ethics of Advanced AI Assistants.**
   Being Human in the Age of AI, Institute of Philosophy, 2024.
   Conference on the Ethics of Generative AI & Conversational Agents, LMU Munich, 2024.

8. **Educational Impact of Advanced AI Assistants.**
   AI in Education Workshop, University of Bristol, 2024.

9. **Autonomy, Advice-Giving, and AI Assistants.**
   Autonomy in an AI World, Imperial College London, 2024.

10. **Engaging Google Teams Through Moral Imagination,** *with Amanda McCroskery, Ben Zevenbergen, et al.*
    Paris Conference on AI & Digital Ethics, 2023.
    NeurIPS Workshop on Cultures in AI/AI in Culture, 2022.

11. **Panel: Fairness in Machine Learning,** *with Sam Corbett-Davies and Susan J. Brison.*
    Neukom Institute for Computational Science, Dartmouth College, 2022.
12. **Algorithmic Bias, Generalist Models, and Clinical Medicine.**
    Workshop on the Ethics of Influence, University of Oxford, 2022.
13. **The Ethics of Autonomous Vehicles.**
    Philosophy and AI Seminar, Meta, 2022.
    Center for Human-Compatible AI, UC Berkeley, 2021.
14. **Enabling Fairness in Healthcare through Machine Learning.**
    Stereotyping and Medical AI Colloqium, King's College London, 2021.
15. **Digital Manipulation and Mental Integrity,** *with Christopher Burr.*
    Workshop on the Philosophy of Online Manipulation, 2021.
16. **A Dilemma for Reasons Additivity**
    Political Theory Workshop, Stanford University, 2021.
17. **On Algorithmic Fairness in Medical Practice,** *with Thomas Grote.*
    Workshop on Algorithmic Fairness, University of Copenhagen, 2020.
18. **Decision Support Systems and Clinical Reasoning,** *with Rune Nyrup.*
    Leverhulme Centre for the Future of Intelligence, University of Cambridge, 2020.
    Philosophy of Medical AI Workshop, University of Tübingen, 2020.
    Issues in Explainable AI, Saarland University, 2019.
19. **Proper Functions: Etiology Without Typehood,** *with Niall Paterson.*
    9th Philosophy of Biology and Cognitive Sciences Workshop, University of the Basque Country, 2019.
    British Society for the Philosophy of Science, University of Durham, 2019.
20. **Object Classification, Autonomous Vehicles, and the Reasonable Belief Standrd**
    2nd CEPPA Graduate Conference on Moral and Political Philosophy, University of St Andrews, 2019.
    Automated Vehicles Symposium, Orlando FL, 2019.
    Research Institute for Ethics and Law, Swansea University, 2019.
    iCog5: Approaches to Higher Cognitive Function, University of Reading, 2019.
    Smart Cities Workshop, University of Bristol, 2019.
21. **Uncertainty About Persons in Autonomous Vehicle Collisions**
    Automated Vehicles Symposium, San Francisco, CA, 2018.
22. **Panel: Ethical and Social Implications of Autonomous Vehicles,** *with Nicholas G. Evans, Noah Goodall, Ryan Jenkins, and Katherine Evans.*
    Automated Vehicles Symposium, San Francisco, CA, 2018.
23. **Against Leben's Rawlsian Algorithm for Autonomous Vehicles**
    Philosophy and Theory of Artificial Intelligence, University of Leeds, 2017.
    Artificial Ethics Symposium, University of Southampton, 2017.

24. **Reliability Weighted Belief Revision in Peer Disagreements,** *with Nemo D'Qrill.*
    Graduate Conference on Epistemology of Disagreement, University of Tartu, 2017.
    9th European Congress of Analytic Philosophy, LMU Munich, 2017.

25. **Blame in Autonomous Vehicle Collisions**
    10th Symposium on Computing and Philosophy, Annual Convention of the Society for the Study of Artificial Intelligence and the Simulation of Behaviour, University of Bath, 2017.

Service

Organising Committee, *ICLR Workshop on Human-AI Coevolution*. 2025
Organising Committee, *Paris Conference on AI & Digital Ethics*. 2024
Program Committee, AI Alignment Track, *AAAI Conference on Artificial Intelligence*. 2024
Program Committee, FAccT, *Association for Computing Machinary*. 2022-2023
Program Committee, Explanatory AI: Between Ethics and Epistemology, *TU Delft*. 2021-2022
Program Committee, Bias and Discrimination in Algorithmic Decisions, *Leibniz University, Hannover*. 2021-2022
Program Committee, Philosophy After AI Symposium, *AISB*. 2019-2020
President, *British Postgraduate Philosophical Association*. 2018-2019
Executive Committee Member, *British Philosophical Association*. 2018-2019

Teaching

Ethics, Law, and Politics of Artificial Intelligence, Primary Instructor, *Stanford University*. 2020-2021
Political Philosophy, Seminar Tutor, *University of Bristol*. 2019-2020
Research Methods for Social Science, Seminar Tutor and Guest Lecturer, *University of Bristol*. 2018-2019
Formal Logic, Seminar Tutor, *University of Bristol*. 2017-2018

Ad-Hoc Teaching
Assessing Consciousness in LLMs, Guest Lecture, *Hong Kong University*. 2025
Fairness in Machine Learning, Guest Lecture, *Princeton University*. 2024
Do LLMs have credences?, Guest Lecture, *Drake University*, with Winnie Street. 2024
Moral Imagination for Responsible Inovation, Guest Lecture, *LMU Munich*, with Benjamin Lange. 2023
Fairness in Machine Learning, Guest Lecture, *Princeton University*. 2023
Black Boxes and Explainability, Guest Lecture, *Chinese University of Hong Kong*. 2021
Ethics of Medical AI, Guest Lecture, *Utrecht University*. 2020

Awards

PhD Studentship, *Arts and Humanities Research Council*, c.a. £55,000. 2017-2020
Awarded through the South, West and Wales Doctoral Training Partnership.
Andrea Mannu Prize, *London School of Economics*. 2016
Awarded for joint-best performance across all undergraduate philosophy programs (joint-1st/51).

Research Students

Lize Alberts, DPhil Candidate in Computer Science, *University of Oxford.* 2023
Primary supervisor for 3 month Student Researcher engagement at Google Research.

Public Policy

Law Commission of England and Wales and Scottish Law Commission. 2021
Consulted for the 2022 joint report on 'Autonomous Vehicles.' [Link]

Wider Engagement

Fireside Chat: AI Cognition and Consciousness, *with Henry Shevlin*. 2025
Homerton College, University of Cambridge.
Panel: Industry-Academia Collaboration in Responsible AI, *with Murray Shanahan & Winnie Street*. 2024
Leverhulme Centre for the Future of Intelligence, University of Cambridge.
Talk: Thinking about AI Consciousness, *with Winnie Street*. 2024
LSE Philosophy Society, London School of Economics.
Panel: AI, Tech and Humanities Careers, *with Harriet Walker, Dora Szabo & Genevieve Liveley*. 2024
South, West and Wales Doctoral Training Partnership.
Panel: Careers in Philosophy, *with Leticia García Martínez, Hamza King, Rachel Ghaw & Jefferson Courtney*. 2024
Department of Philosophy, Logic, and Scientific Method, London School of Economics.
Panel: Careers in AI Ethics and Policy, *with Risto Uuk & Ashyana-Jasmine Kachra*. 2024
London School of Economics.
Podcast: Could AI Undermine Informed Consent?, *with Rune Nyrup & Reid Blackman*. 2024
Ethical Machines Podcast.
Panel: Spotlight on Ethical AI, *with Mandeep Soor, Ismael Garcia & Timothy Wu*. 2023
London School of Economics.
Talk: Ethics and Large Language Models. 2023
Penningtons Manches Cooper & Ethical Reading.
Talk: Practical Ethics in Silicon Valley. 2022
Uehiro Centre for Practical Ethics, University of Oxford.
Talk: Teaching AI Ethics: Lessons from Stanford and Google. 2022
National Symposium on Developing Socially Responsible STEM Professionals, City, University of London.
Talk: Practical Ethics in Tech. 2021
Let's Phi: Beyond Academic Philosophy.
Talk: AI Ethics and the Long Term Future. 2021
CreAItivity Workshop (for underprivileged high school students).
Talk: Philosophy Careers Outside Academia. 2021
Department of Philosophy, University of Bristol.
Talk: Medical Reasoning in the Age of Artificial Intelligence. 2020
Milton Keynes Artificial Intelligence Expert Forum.

Panel: Artificial Intelligence in Healthcare, *with Kourosh Saeb-Parsy, Slawomir Nasuto & Weizi Vicky Li*. 2019
Austin Vita & Ethical Reading Public Event on AI in Healthcare.

Talk: The Ethics of Automated Vehicles. 2018
St Edward's School, Oxford.

Peer Review

Referee for *Nature*, *Nature Machine Intelligence*, *Philosophical Studies*, *British Journal for the Philosophy of Science*, *Journal of Applied Philosophy*, *Utilitas*, *Journal of Ethics*, *Synthese*, *Journal of Medical Ethics*, *Ethics and Information Technology*, *Science and Engineering Ethics*, *Philosophy and Technology*, *Res Publica*, *AI and Society*, *IEEE Transactions on Technology and Society*, *IEEE Intelligent Transportation Systems Transactions*, *IEEE Access*, and *Virtual Reality*. I have also refereed for ACM FAccT, NeurIPS, ICLR, and AAAI/ACM Conference on AI, Ethics and Society, alongside the MIT Press and the John Templeton Foundation.