

Geoff Keeling

📍 London, United Kingdom

in /in/geoff-keeling

🔗 geoffkeeling.github.io

✉ geoffkeelingbhyat@gmail.com

Employment

Senior Research Scientist, <i>Google Research</i> .	2023–Present
Bioethicist, <i>Google Health</i> .	2021–2023
Postdoctoral Fellow, Institute for Human-Centered AI & Center for Ethics in Society, <i>Stanford University</i> .	2020–2021
Research Assistant, Leverhulme Centre for the Future of Intelligence, <i>University of Cambridge</i> .	2019–2020

Education

PhD in Philosophy, <i>University of Bristol</i> , No Corrections.	2017–2020
MA in Philosophy, <i>University of Bristol</i> , Distinction.	2016–2017
BSc in Philosophy, Logic, and Scientific Method, <i>London School of Economics</i> , First Class.	2013–2016

Research

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

Papers

1. Gabriel, I. *, Manzini, A. *, **Keeling, G. ***, Hendricks, L. A., Reiser, V., Iqbal, H., Tomasev, N., Ktena, I., Kenton, Z., Rodriguez, M., El-Sayed, S., Brown, S., Akbulut, C., Trask, A., Hughes, E., Bergman, A. S., Shelby, R., Marchal, N., Griffin, C., Mateos-Garcia, J., Weidinger, L., Street, W., Lange, B., Ingerman, A., Lentz, A., Enger, R., Barakat, A., Krakovna, V., Siy, J. O., Kurth-Nelson, Z., McCroskery, A., Bolina, V., Law, H., Shanahan, M., Alberts, L., Balle, B., de Haas, S., Ibitoye, Y., Dafoe, A., Goldberg, B., Kreier, S., Reese, A., Witherspoon, S., Hawkins, Rauh, M., Wallace, D., Franklin, M., Goldstein, J.A., Lehman, J., Klenk, M., Vallor, S., Biles, C., Ringel Morris, M., King, H., 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](#), [MarkTechPost](#), [MSN](#), [Tech Policy Press](#)]
2. **Keeling, G. ***, Street, W. *, Stachaczyk, M., Zakharova, D., Comşa, 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: [The Guardian](#)]
3. **Keeling, G.** & Street, W. (2024). On the attribution of confidence to large language models. *arXiv preprint arXiv:2407.08388v1*.
4. 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*.
5. 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]

6. 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](#)]
7. 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*.
8. Schaeckermann, M., Spitz, T., Pyles, M., Cole-Lewis, H., Wulczyn, E., Pfhof, 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](#)]
9. 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*.
10. **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).
11. 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)
12. **Keeling, G.** (2023). Algorithmic bias, generalist models, and clinical medicine. *AI and Ethics*, 1-12.
13. **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.
14. **Keeling, G.** (2023). A dilemma for reasons additivity. *Economics & Philosophy*, 39(1), 20-42.
15. Grote, T. & **Keeling, G.** (2022). Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24(3), 39.
16. **Keeling, G.** Automated vehicles and the ethics of classification. In: Jenkins, R., Cerný, D., & Hříbek, T. (Eds). (2022). *Autonomous Vehicle Ethics: The Trolley Problem and Beyond* (pp. 41-57). Oxford University Press.
17. Grote, T. & **Keeling, G.** (2022). On algorithmic fairness in medical practice. *Cambridge Quarterly of Healthcare Ethics*, 31(1), 83-94.
18. **Keeling, G.** & Burr, C. Digital manipulation and mental integrity. In: Jongepier, F., & Klenk, M. (2022). *The Philosophy of Online Manipulation* (pp. 253-271). Routledge.
19. **Keeling, G.** & Paterson, N. (2022). Proper functions: etiology without typehood. *Biology & Philosophy*, 37(3), 19.
20. Bommasani, R., Hudson, D.A., Adeli, E., Altman, R., Arora, S., Arx, S.V., Bernstein, M.S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N.S., Chen, A.S., Creel, K.A., Davis, J., Demszky, D., Donahue, C., Doumbouya, M., Durmus, E., Ermon, S., Etchemendy, J., Ethayarajh, K., Fei-Fei, L., Finn, C., Gale, T., Gillespie, L., Goel, K., Goodman, N.D., Grossman, S., Guha, N., Hashimoto, T., Henderson, P., Hewitt, J., Ho, D.E., Hong, J., Hsu, K., Huang, J., Icard, T.F., Jain, S., Jurafsky, D., Kalluri, P., Karamcheti, S., **Keeling, G.**, Khani, F., Khattab, O., Koh, P., Krass, M.S., Krishna, R., Kuditipudi, R., Kumar, A., Ladhak, F., Lee, M., Lee, T., Leskovec, J., Levent, I., Li, X., Li, X., Ma, T., Malik, A., Manning, C.D., Mirchandani, S., Mitchell, E., Munyikwa, Z., Nair, S., Narayan, A., Narayanan, D., Newman, B., Nie, A., Niebles, J., Nilforoshan, H., Nyarko, J.F., Ogut, G., Orr, L.J., Papadimitriou, I., Park, J.S., Piech, C., Portelance, E., Potts, C., Raghunathan, A., Reich, R., Ren, H., Rong, F., Roohani, Y.H., Ruiz, C., Ryan, J., R'è, C., Sadigh, D., Sagawa, S., Santhanam, K., Shih, A., Srinivasan, K.P., Tamkin, A., Taori, R., Thomas, A.W., Tramèr, F., Wang, R.E., Wang, W., Wu, B., Wu, J., Wu, Y., Xie, S.M., Yasunaga, M.,

You, J., Zaharia, M.A., Zhang, M., Zhang, T., Zhang, X., Zhang, Y., Zheng, L., Zhou, K., & Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. [Media: [Forbes](#), [The Economist](#), [VentureBeat](#), [WIRED](#)]

21. **Keeling, G.** (2020). Why trolley problems matter for the ethics of automated vehicles. *Science and Engineering Ethics*, 26(1), 293-307.
22. **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.
23. **Keeling, G.** (2018). Legal necessity, Pareto efficiency & justified killing in autonomous vehicle collisions. *Ethical Theory and Moral Practice*, 21(2), 413-427.
24. **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.
25. **Keeling, G.** (2018). The sensitivity argument against child euthanasia. *Journal of Medical Ethics*, 44(2), 143-144.
26. **Keeling, G.** (2018). Autonomy, nudging and post-truth politics. *Journal of Medical Ethics*, 44(10), 721-722.
27. **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. **Do Large Language Models Have Credences?**
AI and Ethics Workshop, Princeton University, 2024.
Institute for Ethics in Technology, Hamburg University of Technology, 2024.
2. **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.
3. **Educational Impact of Advanced AI Assistants.**
AI in Education Workshop, University of Bristol, 2024.
4. **Autonomy, Advice-Giving, and AI Assistants.**
Autonomy in an AI World, Imperial College London, 2024.
5. **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.
6. **Panel: Fairness in Machine Learning**, with Sam Corbett-Davies and Susan J. Brison.
Neukom Institute for Computational Science, Dartmouth College, 2022.
7. **Algorithmic Bias, Generalist Models, and Clinical Medicine.**
Workshop on the Ethics of Influence, University of Oxford, 2022.
8. **The Ethics of Autonomous Vehicles.**
Philosophy and AI Seminar, Meta, 2022.
Center for Human-Compatible AI, UC Berkeley, 2021.

9. **Enabling Fairness in Healthcare through Machine Learning.**
Stereotyping and Medical AI Colloquium, King's College London, 2021.
10. **Digital Manipulation and Mental Integrity**, *with Christopher Burr.*
Workshop on the Philosophy of Online Manipulation, 2021.
11. **A Dilemma for Reasons Additivity**
Political Theory Workshop, Stanford University, 2021.
12. **On Algorithmic Fairness in Medical Practice**, *with Thomas Grote.*
Workshop on Algorithmic Fairness, University of Copenhagen, 2020.
13. **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.
14. **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.
15. **Object Classification, Autonomous Vehicles, and the Reasonable Belief Standard**
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.
16. **Uncertainty About Persons in Autonomous Vehicle Collisions**
Automated Vehicles Symposium, San Francisco, CA, 2018.
17. **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.
18. **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.
19. **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.
20. **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 , <i>Paris Conference on AI & Digital Ethics</i> .	2024
Program Committee , AI Alignment Track, <i>AAAI Conference on Artificial Intelligence</i> .	2024
Program Committee , FAccT, <i>Association for Computing Machinery</i> .	2022-2023
Program Committee , Explanatory AI: Between Ethics and Epistemology, <i>TU Delft</i> .	2021-2022
Program Committee , Bias and Discrimination in Algorithmic Decisions, <i>Leibniz University, Hannover</i> .	2021-2022
Program Committee , Philosophy After AI Symposium, <i>AISB</i> .	2019-2020
President , <i>British Postgraduate Philosophical Association</i> .	2018-2019
Executive Committee Member , <i>British Philosophical Association</i> .	2018-2019

Teaching

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

Ad-Hoc Teaching

Fairness in Machine Learning , Guest Lecture, <i>Princeton University</i> .	2024
Do LLMs have credences? , Guest Lecture, <i>Drake University</i> , with Winnie Street.	2024
Moral Imagination for Responsible Innovation , Guest Lecture, <i>LMU Munich</i> , with Benjamin Lange.	2023
Fairness in Machine Learning , Guest Lecture, <i>Princeton University</i> .	2023
Black Boxes and Explainability , Guest Lecture, <i>Chinese University of Hong Kong</i> .	2021
Ethics of Medical AI , Guest Lecture, <i>Utrecht University</i> .	2020

Awards

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

Research Students

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

Public Policy

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

Wider Engagement

Panel: Industry-Academia Collaboration in Responsible AI , with <i>Murray Shanahan & Winnie Street</i> . Leverhulme Centre for the Future of Intelligence, University of Cambridge.	2024
Talk: Thinking about AI Consciousness , with <i>Winnie Street</i> . LSE Philosophy Society, London School of Economics.	2024
Panel: AI, Tech and Humanities Careers , with <i>Harriet Walker, Dora Szabo & Genevieve Liveley</i> . South, West and Wales Doctoral Training Partnership.	2024
Panel: Careers in Philosophy , with <i>Leticia García Martínez, Hamza King, Rachel Ghaw & Jefferson Courtney</i> . Department of Philosophy, Logic, and Scientific Method, London School of Economics.	2024
Panel: Careers in AI Ethics and Policy , with <i>Risto Uuk & Asbyana-Jasmine Kachra</i> . London School of Economics.	2024
Podcast: Could AI Undermine Informed Consent? , with <i>Rune Nyrup & Reid Blackman</i> . Ethical Machines Podcast.	2024
Panel: Spotlight on Ethical AI , with <i>Mandeep Soor, Ismael Garcia & Timothy Wu</i> . London School of Economics.	2023
Talk: Ethics and Large Language Models . Penningtons Manches Cooper & Ethical Reading.	2023
Talk: Practical Ethics in Silicon Valley . Uehiro Centre for Practical Ethics, University of Oxford.	2022
Talk: Teaching AI Ethics: Lessons from Stanford and Google . National Symposium on Developing Socially Responsible STEM Professionals, City, University of London.	2022
Talk: Practical Ethics in Tech . Let's Phi: Beyond Academic Philosophy.	2021
Talk: AI Ethics and the Long Term Future . CreAItivity Workshop (for underprivileged high school students).	2021
Talk: Philosophy Careers Outside Academia . Department of Philosophy, University of Bristol.	2021
Talk: Medical Reasoning in the Age of Artificial Intelligence . Milton Keynes Artificial Intelligence Expert Forum.	2020
Panel: Artificial Intelligence in Healthcare , with <i>Kourosh Saeb-Parsy, Slawomir Nasuto & Weizi Vicky Li</i> . Austin Vita & Ethical Reading Public Event on AI in Healthcare.	2019
Talk: The Ethics of Automated Vehicles . St Edward's School, Oxford.	2018

Peer Review

Referee for *Nature*, *Nature Machine Intelligence*, *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 and AAAI/ACM Conference on AI, Ethics and Society, alongside the MIT Press and the John Templeton Foundation.