Geoff Keeling

London, United Kingdom

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

Employment

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

Education

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

Research

Ethics of AI/ML, Moral and Political Philosophy, Bioethics, Philosophy of Biology, 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 News, The Information, Axios, MarkTechPost, MSN News]

2. 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*.

3. 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, Digital Information World]

4. 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.*

5. 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]

6. 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.*

7. 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)

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

9. **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.

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

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

12. **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.

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

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

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

16. 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'e, 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]

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

18. **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.

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

20. **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.
21. **Keeling, G.** (2018). The sensitivity argument against child euthanasia. *Journal of Medical Ethics*, *44*(2), 143-144.
22. **Keeling, G.** (2018). Autonomy, nudging and post-truth politics. *Journal of Medical Ethics*, 44(10), 721-722.
23. **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.
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 Colloqium, King's College London, 2021.
10. **Digital Manipulation and Mental Integrity,** *with Christopher Burr.*
    Workshop on the Philosophy of Online Manipulation, 2021.
11. **Proper Functions: Etiology Without Typehood,** *with Niall Paterson.*
    British Society for the Philosophy of Science, University of Durham, 2019.
12. **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.
13. **Uncertainty About Persons in Autonomous Vehicle Collisions**
    Automated Vehicles Symposium, San Francisco, CA, 2018.

Service

Organising Committee, *Paris Conference on AI & Digital Ethics*. 2024-Present
FAccT Program Committee, *Association for Computing Machinary*. 2018-2019
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
Knowledge and Reality, Exam Marker, *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

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]

Peer Review

Referee for *Nature*, *Nature Machine Intelligence*, *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.