Carina Prunkl

5 Canal Street, Oxford, OX2 6BQ, UK
+44 7546 833734
carina.prunkl@philosophy.ox.ac.uk
www.carinaprunkl.com

Academic Appointments and Affiliations

Research Fellow, Institute for Ethics in AI, University of Oxford 2020–2023
Senior Research Scholar, Future of Humanity Institute, University of Oxford 2018–2020
Research Affiliate, Black Hole Initiative, Harvard University since 2019

Other Appointments

Ethics Advisor, Artificial Intelligence Lab, VU Brussels since 2021
Ethics Advisor, Digital Catapult 2021–2022
Ethics Advisor, "Prediction of radiotherapy side effects using explainable AI for patient communication and treatment modification", Horizon Europe Framework Programme 2022–2026

Education

DPhil Philosophy, University of Oxford 2014–2018
MSt Philosophy of Physics (Distinction), University of Oxford 2013–2014
MSc Physics (equiv. First Class Honours), Freie Universität Berlin 2011–2013
BSc Physics, Freie Universität Berlin 2007-2011

Publications

Journal Publications "LUCID: Exposing Algorithmic Bias through Inverse Design" with C. Mazijn, J. Danckaert, and V. Ginis. *Proceedings of the 37th AAAI Conference on Artificial Intelligence* (forthcoming).

"Epistemic Injustice and Algorithmic Profiling" with S. Milano.*Philosophical Studies* (forthcoming).

"How objective is thermodynamics?" with K. Robertson. *Philosophy of Science* (forthcoming).

"Human Autonomy in the Age of Artificial Intelligence" *Nature Machine Intelligence* (2022) 4.2: 99-101.

"Is there a trade-off between human autonomy and the 'autonomy' of AI systems?" *Philosophy and Theory of Artificial Intelligence 2021*, ed. Müller, C, Springer Cham (2022).

"Institutionalising Ethics in AI through Broader Impact Requirements" with C. Ashurst, M. Anderljung, H. Webb, J. Leike, and A. Dafoe. *Nature Machine Intelligence* (2021) 3:104–110.

"We might be afraid of black-box algorithms" with C. Véliz, M. Phillips-Brown, and T. Lechterman. *Journal of Medical Ethics* (2021) 47.5:339-340.[1]

"Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims" with M. Brundage et al. *Report* (2020), http://www.towardtrustworthyai.com/.

"Beyond Near-and Long-Term: Towards a Clearer Account of Research Priorities in AI Ethics and Society." *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (2020) pp. 138–143.

"On the Equivalence of Thermodynamic and von Neumann Entropy" *Philosophy of Science* (2020), 87, 2:262–280.

"On the Thermodynamical Cost of Some Quantum Interpretations" with C. Timpson. *Studies in History and Philosophy of Modern Physics* (2018) 63:114–122.

"Impulsivity, self-control, and hypnotic suggestibility" with V. Ludwig et al. *Consciousness and Cognition* (2013) 22(2):637–653.

Books

"Entropy" *Cambridge Elements - Philosophy of Physics*, J.O. Weatherall (ed.), Cambridge University Press, (forthcoming 2022).

[1] This is an invited, internally peer-reviewed commentary.

Manuscripts "LUCID–GAN: Conditional Generative Models to Locate Unfairness" with A. Algaba, C. Mazijn, J. Danckaert, and V. Ginis *SSRN 4289597*.

"Artificial Intelligence and Human Autonomy—a philosophical perspective", (journal submission).

"Simulation Intelligence: Towards a New Generation of Scientific Methods" with A. Lavin et al. *arXiv:2112.03235*.

Blog posts

"A Guide to Writing the NeurIPS Impact Statement" (with C. Ashurst, M. Anderljung, J. Leike, Y. Gal, T. Shevlane, A. Dafoe), Medium post, 13/05/2020.

Selected Invited Talks and Conference Contributions

2022

Transatlantic Dialogue on Humanity and AI Regulation (Paris); Blavatnik School of Government; Institute for Advanced Study in Toulouse.

2021

Philosophy of Science meets Machine Learning (Tübingen); Workshop Responsible AI in the Defence Sector (Oxford); 4th Conference on Philosophy and Theory of Artificial Intelligence (Goetheburg); The Oxford Union; Surrey Centre for Law and Philosophy.

2020

AAAI/ACM Conference on AI, Ethics, and Society; Women Leading in AI; University of Michigan; AI Lab, Vrije Universiteit Brussel; European Institute for Participatory Media; Senate, Mexico City; London School of Economics.

Policy Engagement

- United Nations Interregional Crime and Justice Research Institute (UNICRI); Centre for Data Ethics and Innovation, UK; Ministry of Defence, UK; UK2070 Independent Commission, UK; Senate, Mexico; EU Delegation in Russia; Congress, Mexico.

Teaching Experience

Ethics of AI, Lecturer and Tutor

- University of Oxford (MSt Practical Ethics)
- University of Oxford (3rd year philosophy undergraduates) 2021
- Scuola Internazionale Superiore di Studi Avanzati, Trieste (PhD in computer science) 2021
- Thales Akademie, Germany (professionals) 2021
- Said Business School, University of Oxford (professional fellows) 2019
- Oxford Artificial Intelligence Society (undergraduates and graduates) 2019

Governance of AI, Lecturer

- Department of Engineering Sciences, University of Oxford (PhD in Autonomous Intelligent Machines and Systems) since 2019

Other teaching (Oxford): Logic (undergraduate); Advanced Philosophy of Physics (graduate); Philosophy of Science (undergraduate); Quantum Theory and Quantum Computation (undergraduate)

Awards and Funding

Santander-CIDOB '35 under 35 Future Leaders' list 2021
Oxford Vice Chancellor's Fund Award 2017
BSPS Doctoral Scholarship Award 2014–2017
Konrad Adenauer Foundation Scholarship 2008–2013

Languages

German (mother tongue), English (business fluent), French (fluent), Spanish (intermediate)