Carina Prunkl

Institute for Ethics in AI, Oxford, UK
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

Visiting Fellowships
Visiting Postdoctoral Research Fellow, Harvard University Mar–Apr 2019
Visiting Research Fellow, Black Hole Initiative, Harvard University Mar–May 2018
Visiting Fellow, Munich Centre for Mathematical Philosophy, LMU Munich Summer 2017
Visitor, Center for Quantum Technologies, National University of Singapore Summer 2016

Other Appointments

Ethics Advisor, Artificial Intelligence Lab, VU Brussels since 2021
Ethics Advisor, Digital Catapult, UK 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
- Thesis title: The Scope of Thermodynamics
- Advisors: Prof. C. Timpson and Prof. H. Brown
- Additional modules: Political Philosophy, Epistemology

MSt Philosophy of Physics (Distinction), University of Oxford 2013–2014
- Modules: Philosophy of Physics, Philosophy of Mind, Philosophy of Science

MSc Physics (equiv. First Class Honours), Freie Universität Berlin 2011–2013
- Thesis topic: Superactivation in Quantum Information Theory

BSc Physics, Freie Universität Berlin 2007-2011

Publications

Journal Publications

Milano, S. and Prunkl, C. (*forthcoming*). Epistemic Injustice and Algorithmic Profiling. In *Philosophical Studies*.

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

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

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

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

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

Prunkl, C. (2020) On the Equivalence of Thermodynamic and von Neumann Entropy. In *Philosophy of Science*, 87, 2:262-280

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

Ludwig, V., Stelzel, C., Krutiak, H., Prunkl, C., Steimke, R., Paschke, L., Kathmann, N., and Walter, H. (2013) In *Consciousness and Cognition*, 22(2):637–653.

Books and Book Chapters

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

Prunkl, C. (*forthcoming*). Quantum Thermodynamics. In *Quantum Foundations of Statistical Mechanics*, eds. D. Bedingham et al., Oxford University Press.

Manuscripts

Mazijn, C., Danckaert, J., Prunkl, C., and Ginis, V. (*journal submission*) Exposing Algorithmic Bias through Inverse Design.

Lavin et al. (2021) Simulation Intelligence: Towards a New Generation of Scientific Methods, preprint available *arXiv:2112.03235*.

Prunkl, C. and Timpson, C. (*revise and resubmit, BJPS*) Black Hole Entropy is Thermodynamic Entropy, preprint available *arxiv:1903.06276*.

Blog posts and popular Articles

"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

"Endlich Unendlich - Auf der Suche nach dem ewigen Leben", *SHIFT*, 4:14–19, 2016

"Das Schummeln der Lämmer - Von kleinen Lügen und großen Konsequenzen", *SHIFT*, 1:42–46, 2013

Selected Invited Talks

Limitations of risk-based regulation
HEC Paris, Transatlantic Dialogue on Humanity and AI Regulation *2022*

Human Autonomy and Artificial Intelligence
Institute for Advanced Study in Toulouse, Research Seminar *2022*

Principles for Autonomous Systems
Blavatnik School of Government, CDEI, and MoD, Workshop: Autonomy and the Defence Sector *2021*

Governance from within
Tübingen University, Philosophy of Science meets Machine Learning *2021*

Is there a trade-off between human autonomy and system autonomy?
University of Goetheburg, 4th Conference on Philosophy and Theory of Artificial Intelligence *2021*

Staying in Charge: Human Autonomy and Artificial Intelligence
The Oxford Union, AI: Bridging Technology and Governance *2021*

Algorithmic Profiling and Epistemic Injustice
Surrey Centre for Law and Philosophy, Symposium on Algorithms and Ethics *2021*

Accountability mechanisms for responsible AI development
Women Leading in AI Webinar *2020*

How Anthropocentric is Thermodynamics?
University of Michigan, International Postdoc Forum for the Philosophy of Science *2020*

AI and Human Autonomy
Artificial Intelligence Lab, Vrije Universiteit Brussel *2020*

Echo Chambers and Decision-Making
European Institute for Participatory Media, Workshop on Reflective AI *2020*

AI Ethics and Governance
Senate, Mexico City, Un acercamiento a la Inteligencia Artificial *2020*

How anthropocentric is thermodynamics?
London School of Economics, LSE Sigma Club *2019*

Regulating AI? – challenges and opportunities for the responsible development of new tech
Moscow, EMERTECH conference, EU Delegation in Russia *2019*

Future Technology Workshop
UK2070 Commission, Discussant *2019*

Boltzmann Brains Simulations - Rethinking the Skeptical Hypothesis
University of Bonn, Philosophy of Physics Seminar *2019*

Black Hole Entropy is Thermodynamic Entropy
London School of Economics, LSE Sigma Club *2018*

Black Hole Entropy — how much information do we need?
Harvard University, The Black Hole Initiative Colloquium *2017*

Selected Conference Contributions

Philosophy and Theory of Artificial Intelligence
Goetheburg, Human Autonomy and Artificial Intelligence *2021*

AAAI/ACM Conference on AI, Ethics, and Society
New York, Beyond Near- and Long-Term: Research Priorities in AI Ethics and Society *2020*

European Philosophy of Science Association Conference
University of Geneva, Black Holes and Information *2019*

British Society for the Philosophy of Science Annual Conference (BSPS)
Oxford, Symposium on Black Holes: Entropy and System Size *2018*

Conference on the Second Law of Thermodynamics
LMU Munich, "Thermodynamics Without Observers" *2017*

Black Forest Summerschool in Philosophy of Physics
Saig, Germany, "Black Hole Entropy is Entropy (and not Information)" *2017*

Philosophy of Science Association (PSA)
Atlanta, "A Tale of Two Entropies - Defending the von Neumann Entropy" *2016*

British Society for the Philosophy of Science Annual Conference
Cardiff, "Are Some Quantum Interpretations Truly Hotter Than Others?" *2016*

Workshop Metaphysics and Philosophy of Physics
University of Bristol, Invited Discussant *2016*

Policy Engagement

- United Nations Interregional Crime and Justice Research Institute (UNICRI)
- Centre for Data Ethics and Innovation, Dep. for Digital, Culture, Media & Sport, UK
- Ministry of Defence, UK
- UK2070 Independent Commission, UK
- Senate, Mexico
- Ministry of Economics, Mexico
- EU Delegation in Russia

Teaching Experience

Ethics of AI, Lecturer and Tutor
- Philosophy Faculty, University of Oxford (graduate) — MT 2021

Epistemic Injustice, Undergraduate Thesis Supervisor
- Philosophy Faculty, University of Oxford (undergraduate) — MT 2021

Governance of AI, Lecturer
- Department of Engineering Sciences, University of Oxford (graduate) — since 2019

Ethics and Social Implications of AI, Lecturer
- Said Business School, University of Oxford (professional fellows) — MT 2019
- Oxford Artificial Intelligence Society (undergraduate and graduate) — HT 2019
- Scuola Internazionale Superiore di Studi Avanzati, Trieste (Masterclass) — Jan 2021

Advanced Philosophy of Physics, Lecturer
- Faculty of Philosophy, University of Oxford (graduate) — MT 2019, 2021

Introduction to Logic, Teaching Assistant
- Hertford College (undergraduate)

Philosophy of Science, Tutor
- Balliol College, Worcester College (undergraduate) — 2015–2016

Quantum Theory and Quantum Computers, Teaching Assistant
- Mathematical Institute, University of Oxford (undergraduate) — HT 2014

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

Academic Service

Member, Humanities Cultural Programme Steering Group, University of Oxford — 2022
Programme Committee, ACM Conference on Fairness, Accountability, and Transparency — 2022
Convenor, Ethics and AI Lunch Seminar, University of Oxford — 2022
Programme Committee, AAAI/AIES Conference — 2021
Board Member, Oxford AI Society — since 2019
Expert Panel, Mentor, A.I. Impact Weekend, Oxford Foundry, University of Oxford — 2019
Undergraduate Admission Inteviews, Oriel, Balliol and Brasenose College — 2016–2019
Organising Committee — 2016–2017
DPhil Seminar, University of Oxford
Founder, Philosophy of Physics Graduate Lunch Seminar (POP-Grunch) — since 2015
Organising Committee — 2015

Conference in honour of Prof. Harvery Brown's 65th Birthday
Referee, British Journal for the Philosophy of Science, Philosophy of Science,
Studies in History,and Philosophy of Modern Physics, Analysis, Review of Social Economy,
Nature Machine Intelligence, Synthese, AI and Society since 2015