Professor Seth Lazar
School of Philosophy
Machine Intelligence and Normative Theory (MINT) Lab

seth@mintresearch.org
sethlazar.org
mintresearch.org

Employment

2026- Incoming Professor. School of Government and Policy, Johns Hopkins University
2025 Visiting Faculty Researcher. Google (0.2FTE, six-month appointment)
22-25 ARC Future Fellow. School of Philosophy, RSSS, ANU
20-26 Professor. School of Philosophy, RSSS, ANU
18-21 Project Leader. Humanising Machine Intelligence Grand Challenge, ANU
17-19 Head of School. School of Philosophy, RSSS, ANU
17-19 Associate Professor. School of Philosophy, RSSS, ANU
2015- Senior Research Fellow. School of Philosophy, RSSS, ANU
11-14 Continuing Research Fellow. School of Philosophy, RSSS, ANU
09-11 Research Associate, Institute for Ethics, Law, and Armed Conflict (ELAC), University of Oxford
07-09 Retained Lecturer in Political Theory, Pembroke College, Oxford

Education

06-09 D.Phil. Politics (Political Theory), Department of Politics, Oxford
'War and Associative Duties'. Supervisor Henry Shue, examined by Jeff McMahan and David Rodin
04-06 M.Phil. Politics (Political Theory), Department of Politics, Oxford
Distinction. 'A Critical Analysis of Corrective Justice', supervised by David Miller
99-02 BA (Hons) English Language and Literature, Wadham College, Oxford
First class honours.

Fellowships, Honours

24-25 Visiting Professor, School of Philosophy, Hong Kong University
24-25 Senior AI Advisor, Knight First Amendment Institute, Columbia University
2024- Non-Resident Fellow at the Carnegie Endowment for International Peace
2023 Tanner Lectures on AI and Human Values, Human-Centred AI Institute and McCoy Center for Ethics in Society, Stanford University
2022 Mala and Solomon Kamm Lecture in Ethics, Safra Centre for Ethics, Harvard University
22-26 ARC Future Fellow. School of Philosophy, RSSS, ANU
2021- Distinguished Research Fellow, Institute for Ethics in AI and Faculty of Philosophy, University of Oxford (honorary position)
2021 Member of 14-person US National Academies of Science, Engineering and Medicine Study Committee on 'Ethics and Governance of Responsible Computing Research', led by Professor Barbara Grosz (Harvard)
2019 Vice-Chancellor's Award for Excellence in Research, ANU
2016 Academy of Social Sciences of Australia Early Career Researcher Commendation (Panel D: History and Philosophy)
13-15 ARC Discovery Early Career Research Award Fellow. School of Philosophy, RSSS, ANU
2012 Carnegie Ethics Council Global Ethics Fellowship
2011 Visiting fellow at programme on Sovereignty, Global Justice, and the Ethics of War, Institute of Advanced Studies (IAS), Hebrew University of Jerusalem
2011 American Philosophical Association Frank Chapman Sharp memorial prize for the best unpublished monograph on the philosophy of war and peace
2009 *Res Publica* Postgraduate Essay Prize for 2008

2008 Society for Applied Philosophy Annual Conference Postgraduate Essay Prize
2007 Social Science Division Teaching Excellence Award (Category A)
02-03 Harvard University, Frank Knox Fellowship
00-02 Wadham College, Schools Prize and Junior Scholarship

Research Publications

Monographs in Preparation

2026 *The Algorithmic City: Power, Justice, and AI.* Stanford Tanner Lectures on AI and Human Values (edited by Rob Reich, with responses by Marion Fourcade, Arvind Narayanan, Renee Jorgensen, and Josh Cohen).
Under contract with Oxford University Press, final MS submitted

Monographs

2016 *Sparing Civilians*. Oxford: Oxford University Press.

Edited Volumes and Symposia

2025 AI and Democratic Freedoms. *Knight First Amendment Institute Symposium* (with Katy Glenn Bass). https://knightcolumbia.org/events/artificial-intelligence-and-democratic-freedoms

2023 Normative Theory and AI. *Philosophical Studies* (guest editor with Claire Benn, Todd Karhu, and Pamela Robinson), https://link.springer.com/collections/ccgjbejhce

2022 The Political Philosophy of Data and AI. *Canadian Journal of Philosophy* (guest editor with Kate Vredenburgh and Annette Zimmermann), https://www.cambridge.org/core/journals/canadian-journal-of-philosophy/article/political-philosophy-of-data-and-ai/B7C38B3D585739080CC94EC48A6E6D7C

2021 Proceedings of the 2021 AAAI/ACM Artificial Intelligence, Ethics, and Society Conference (Co-Editor, with Marion Fourcade, Ben Kuipers, Deirdre Mulligan). https://dl.acm.org/doi/proceedings/10.1145/3461702

2020 Topical Collection on Norms for Risk. *Synthese* (guest editor, with Alan Hájek and lead editor Renee Jorgensen). https://link.springer.com/collections/gbchegjeba

2018 *The Oxford Handbook of Ethics of War*. Oxford: Oxford University Press (Co-Editor, with Helen Frowe), https://academic.oup.com/edited-volume/28369

2017 Symposium on Ethics and Decision Theory. *Ethics* (guest editor), 127/3 https://www.jstor.org/stable/e26541029

2014 *The Morality of Defensive War*. Oxford: Oxford University Press. (Co-Editor, with Cécile Fabre), https://academic.oup.com/book/8445

2011 Symposium on Jeff McMahan's *Killing in War*. *Ethics* (guest editor), 122/1, https://www.jstor.org/stable/10.1086/662055

Pre-Prints

2025 'Military AI Cyber Agents (MAICAs) Constitute a Global Threat to Critical Infrastructure', with Tim Dubber (lead author), https://arxiv.org/abs/2506.12094 (accepted to *Regulatable ML Workshop*, NeurIPS 2025)

Published Articles

2026 'Using LLMs to Advance Democratic Values', with Lorenzo Manuali, *Minds and Machines*, (accepted January 2026) https://arxiv.org/abs/2410.08418

'Resource Rational Contractualism Should Guide AI Alignment', with Sydney Levine (lead author), Matija Franklin, Tan Zhi-Xuan, Secil Yanik Guyot, Lionel Wong, Daniel Kilov, Yejin Choi, Joshua B. Tenenbaum, Noah Goodman, Seth Lazar, Iason Gabriel, International Association for Safe and Ethical AI Conference (IASEAI), Paris 2026, https://www.arxiv.org/abs/2506.17434

'Discerning What Matters: A Multi-Dimensional Assessment of Moral Competence in LLMs', with Daniel Kilov, Caroline Hendy, Secil Yanik Guyot, and Aaron J. Snoswell, https://www.arxiv.org/abs/2506.13082 IASEAI Paris 2026 (also accepted to *Foundations of Reasoning in Language Models Workshop*, NeurIPS 2025)

'Beyond Verdicts: Evaluating Language Model Moral Competence', with Aaron J. Snoswell and Daniel Kilov, *Association for the Advancement of Artificial Intelligence Conference (AI Alignment Track),* https://philpapers.org/rec/SNOBVE

2025 'The AI Power Disparity Index: Toward a Compound Measure of AI Actors' Power to Shape the AI Ecosystem', with Rachel Kim and Blaine Kuehnert (lead authors), Ranjit Singh, and Hoda Heidari *Knight First Amendment Institute,* September 2025, https://knightcolumbia.org/content/the-ai-power-disparity-index-toward-a-compound-measure-of-ai-actors-power-to-shape-the-ai-ecosystem (also published in ACM/AAAI AI, Ethics and Society Conference, 2025)

'On the Moral Case for Using Language Model Agents for Recommendation', with Luke Thorburn, Tian Jin, and Luca Belli, *Inquiry*, https://www.tandfonline.com/doi/full/10.1080/0020174X.2025.2515579?src=exp-la

'AI Agents and Democratic Resilience', with Mariano-Florentino Cuéllar, *Knight First Amendment Institute,* September 2025, https://knightcolumbia.org/content/ai-agents-and-democratic-resilience

'Anticipatory AI Ethics', *Knight First Amendment Institute*, April 2025, https://knightcolumbia.org/content/anticipatory-ai-ethics

'Infrastructure for AI Agents', with Alan Chan (lead author), Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K Hadfield, Markus Anderljung, *Transactions of Machine Learning Research*, https://arxiv.org/abs/2501.10114

'Position: Build Agent Advocates, not Platform Agents', with Sayash Kapoor and Noam Kolt, *International Conference on Machine Learning*, https://arxiv.org/abs/2505.04345

'Using LLMs to Enhance Democracy' with Lorenzo Manuali, *ACM Fairness, Accountability, and Transparency Conference* (non-archival), https://arxiv.org/abs/2410.08418

'Governing the Algorithmic City', *Philosophy & Public Affairs*, 53/2, 102-168 https://onlinelibrary.wiley.com/doi/10.1111/papa.12279?af=R

2024 'Position: On the Societal Impact of Open Foundation Models', with lead authors Rishi Bommasani, Sayash Kapoor, et al, *International Conference on Machine Learning*, https://arxiv.org/abs/2403.07918

'Attention, Moral Skill, and Algorithmic Recommendation', with Nick Schuster, *Philosophical Studies*, 182/1, 159-184, https://link.springer.com/article/10.1007/s11098-023-02083-6

'On the Site of Predictive Justice', with Jake Stone, *Noûs*, 58/3, 730-754, https://onlinelibrary.wiley.com/doi/10.1111/nous.12477

'Legitimacy, Authority, and Democratic Duties of Explanation', *Oxford Studies in Political Philosophy*, Volume 10, 28-56, https://academic.oup.com/book/56337/chapter-abstract/445461225

'Frontier AI Ethics', *Aeon*, https://aeon.co/essays/can-philosophy-help-us-get-a-grip-on-the-consequences-of-ai and https://arxiv.org/abs/2404.06750

'Can Democracy Survive Artificial General Intelligence?', with Alex Pascal, *Tech Policy Press*, https://www.techpolicy.press/can-democracy-survive-artificial-general-intelligence/

2023 'Communicative Justice and the Distribution of Attention', *Knight First Amendment Institute,* https://knightcolumbia.org/content/communicative-justice-and-the-distribution-of-attention

'On the Site of Predictive Justice', with Jake Stone, *ACM Fairness, Accountability, and Transparency Conference* (non-archival) (published in Noûs, above)

'AI Safety on Whose Terms?' with Alondra Nelson, *Science*, https://www.science.org/doi/10.1126/science.adi8982

2022 'Supererogation and Optimisation', with Christian Barry, *Australasian Journal of Philosophy*, 102/1, 21-36, https://www.tandfonline.com/doi/abs/10.1080/00048402.2022.2074066

'What's Wrong with Automated Influence', with Claire Benn, *Canadian Journal of Philosophy*, 52/1, 125-148, https://philarchive.org/rec/BENWWW-3

2021 'Deontological Decision Theory and Lesser-Evil Options', with Peter A. Graham, *Synthese*, 198, 6889-6916, https://www.jstor.org/stable/27293779

2020 'Duty and Doubt', *Journal of Practical Ethics*, 8/1, 28-55, https://www.jpe.ox.ac.uk/wp-content/uploads/2020/06/JPE0053-Lazarr.pdf

'Should I Use that Rating Factor? A Philosophical Approach to an Old Problem' with Chris Dolman (lead author), Tiberio Caetano, and Dimitri Semenovich, *2020 All Actuaries Summit,* https://hmi.anu.edu.au/ourwork/should-i-use-that-rating-factor-a-philosophical-approach-to-an-old-problem

2019 'Deontological Decision Theory and the Grounds of Subjective Permissibility', *Oxford Studies in Normative Ethics*, 9, https://www.jstor.org/stable/27293779

'Self-Ownership and Agent-Centred Options', *Social Philosophy and Policy*, 36/2, 36-50, https://philpapers.org/rec/LAZSAA-2

'Accommodating Options', *Pacific Philosophical Quarterly*, 100/1, 233-255, https://philarchive.org/rec/LAZAO

'Moral Status and Agent-Centred Options', *Utilitas*, 31/1, 83-105, https://philarchive.org/rec/LAZMSA

'Axiological Absolutism and Risk', with Chad Lee-Stronach, *Noûs*, 53/1, 97-113, https://onlinelibrary.wiley.com/doi/am-pdf/10.1111/nous.12210

2018 'Limited Aggregation and Risk', *Philosophy & Public Affairs*, 46/2, 117-159, https://philarchive.org/archive/LAZLAA

'Moral Sunk Costs', *Philosophical Quarterly*, 68/273, 841-861, https://academic.oup.com/pq/article-abstract/68/273/841/5051223

'Strengthening Moral Distinction' (response to symposium on *Sparing Civilians*), *Law and Philosophy*, 37/3, 327-349, https://www.jstor.org/stable/44980942

'In Dubious Battle: Uncertainty and the Ethics of Killing', *Philosophical Studies*, 175/4, 859-883, https://www.jstor.org/stable/45094242

2017 'Risky Killing: How Risks Worsen Violations of Objective Rights', *Journal of Moral Philosophy*, 16/1, 1-26, https://philpapers.org/rec/LAZRK

'Deontological Decision Theory and Agent-Centred Options', *Ethics*, 127/3, 579-609, https://www.journals.uchicago.edu/doi/abs/10.1086/690069

'Response: Limiting Defensive Rights', *Journal of Applied Philosophy*, 34/1, 19-23, https://philpapers.org/rec/LAZRLD

'Proxy Battles in Just War Theory: Jus in Bello, the Site of Justice, and Feasibility Constraints', with Laura Valentini, *Oxford Studies in Political Philosophy*, volume 3, https://philarchive.org/archive/LAZPBI

'Evaluating the Revisionist Critique of Just War Theory', *Daedalus*, 146/1, 113-124, https://direct.mit.edu/daed/article/146/1/113/27135/Evaluating-the-Revisionist-Critique-of-Just-War

'Just War Theory: Revisionists Vs Traditionalists', *Annual Review of Political Science*, 20, 37-54, https://www.annualreviews.org/content/journals/10.1146/annurev-polisci-060314-112706

'Anton's Game: Deontological Decision Theory for an Iterated Decision Problem', *Utilitas*, 29/1, 88-109, https://philpapers.org/rec/LAZAGD

2016 'Complicity, Collectives, and Killing in War', *Law and Philosophy*, 35/4, 365-389, https://www.jstor.org/stable/44980900

'The Justification of Associative Duties', *Journal of Moral Philosophy*, 13/1, 28-55, https://philarchive.org/rec/LAZTJO

'Authorization and the Morality of War', *Australasian Journal of Philosophy*, 94/2, 211-226, https://philpapers.org/rec/LAZAAT-2

2015 'Risky Killing and the Ethics of War', *Ethics*, 126/1, 91-117, https://www.jstor.org/stable/10.1086/682191?seq=1

'Authority, Oaths, Contracts, and Uncertainty in War', *Thought*, 4/1, 52–58, https://www.pdcnet.org/collection/fshow?id=tht_2015_0004_0001_0052_0058&pdfname=tht_2015_0004_0001_0055_0061.pdf&file_type=pdf

2014 'Necessity and Noncombatant Immunity', *Review of International Studies*, 40/1, 53-76, https://www.cambridge.org/core/journals/review-of-international-studies/article/abs/necessity-and-noncombatant-immunity/55A56B582C7EEE5C4D81595FAB83A7D6

2013 'Associative Duties and the Ethics of Killing in War', *Journal of Practical Ethics*, 1/1, 3-48, https://www.jpe.ox.ac.uk/wp-content/uploads/2013/06/JPE0001-Lazar.pdf

2012 'Necessity in Self-Defense and War', *Philosophy & Public Affairs*, 40/1, 3-44, https://www.jstor.org/stable/23261274

2010 'A Liberal Defence of (Some) Duties to Compatriots', *Journal of Applied Philosophy*, 27/3, 246-257, https://www.jstor.org/stable/24356040

'The Responsibility Dilemma for *Killing in War*: A Review Essay', *Philosophy & Public Affairs*, 38/2, 180-213, https://philosophy.rutgers.edu/joomlatools-files/docman-files/Philosophy%20&%20Public%20Affairs.pdf

2009 'Responsibility, Risk, and Killing in Self-Defense', *Ethics*, 119/4, 699-728, https://www.jstor.org/stable/10.1086/605727

'Debate: Do Associative Duties Really Not Matter?', *Journal of Political Philosophy*, 17/1, 90-101, https://philarchive.org/rec/LAZDDA-2

'The Nature and Disvalue of Injury', *Res Publica*, 15/3, 289-304, https://philarchive.org/rec/LAZTNA

2008 'Corrective Justice and the Possibility of Rectification', *Ethical Theory and Moral Practice*, 11/4, 355–68, https://philarchive.org/rec/LAZCJA-2

Chapters in Peer-Reviewed Edited Volumes

2024 'Automatic Authorities: Power and AI', *Collaborative Intelligence: How Humans and AI are Transforming our World*, Arathi Sethumadhavan and Mira Lane (eds.), Cambridge: MIT Press, https://direct.mit.edu/books/edited-volume/5886/chapter-abstract/5154049/Automatic-Authorities-Power-and-AI?redirectedFrom=fulltext

2022 'Power and AI', *The Oxford Handbook of AI Governance*, Johannes Himmelreich et al. (eds.), New York: Oxford University Press, https://academic.oup.com/edited-volume/41989/chapter-abstract/355437737?redirectedFrom=fulltext

2018 'The Ethics of War', (with Helen Frowe) *The Oxford Handbook of Ethics of War*, Seth Lazar and Helen Frowe (eds.), New York: Oxford University Press, https://academic.oup.com/edited-volume/28369

'Method in the Morality of War', *The Oxford Handbook of Ethics of War*, Seth Lazar and Helen Frowe (eds.), New York: Oxford University Press, https://academic.oup.com/edited-volume/28369/chapter-abstract/215247461?redirectedFrom=fulltext

2016 'War', *Stanford Encyclopedia of Philosophy*, Zalta (ed.), Spring 2016 edition, http://plato.stanford.edu/entries/war/

'War's Endings and the Structure of Just War Theory', *The Ethics of War*, Sam Rickless and Saba Bazargan (eds.), New York: Oxford University Press, https://academic.oup.com/book/7553/chapter-abstract/152537951?redirectedFrom=fulltext

'Travel, Friends, and Killing', in *Philosophers Take on the World*, Edmonds (ed.), Oxford: Oxford University Press, 25-27, https://philpapers.org/rec/LAZTFA-2 https://global.oup.com/academic/product/philosophers-take-on-the-world-9780198822639?cc=au&lang=en&

'The Associativist Account of the Ethics of War', *Global Political Theory,* David Held and Pietro Maffettone (eds.), Cambridge: Polity, 158-179, https://catalogue.nla.gov.au/catalog/7246646

'Liability and the Ethics of War: A Reply to Strawser and McMahan', *The Ethics of Self-Defence*, Coons and Weber (ed.), New York: Oxford University Press, 292-304, https://academic.oup.com/book/27662/chapter-abstract/197785102?redirectedFrom=fulltext

2014 'National Defence, Self-Defence, and the Problem of Political Aggression', in *The Morality of Defensive War*, Lazar and Fabre (eds.), Oxford: Oxford University Press, 9-37, https://philarchive.org/rec/LAZNDS

2013 'War', *International Encyclopaedia of Ethics*, Lafollette (ed.), Wiley-Blackwell, https://www.academia.edu/5281771/_War_International_Encyclopaedia_of_Ethics_Lafollette_ed_Wiley_Blackwell

'Just War Theory', *Oxford Companion to Comparative Politics*, Joel Krieger (ed.), Oxford University Press.

2012 'Scepticism about *Jus Post Bellum*', *Morality, Jus Post Bellum, and International Law* Larry May, Andrew Forcehimes (eds.), Cambridge: CUP, 204-22, https://philarchive.org/rec/LAZSAJ

'The Morality and Law of War', *Routledge Companion to Philosophy of Law*, Andrei Marmor (ed.), London: Routledge, 364-79, https://philarchive.org/archive/LAZTMA

National Academy Reports

2022 *Fostering Responsible Computing Research: Foundations and Practices*, A Consensus Study Report of the National Academies of Sciences, Engineering Medicine, National Academies Press. Lead Author Barbara Grosz (Harvard). Report commissioned by the NSF.

Other Writing and Interviews

2024 'Seth Lazar: Normative Philosophy of Computing' *The Gradient*, podcast interview, https://thegradientpub.substack.com/p/seth-lazar-normative-philosophy-of-computing

'The Rise and Fall (and Rise Again) of the First AI Agent Millionaire', *Tech Policy Press*, https://www.techpolicy.press/the-rise-and-fall-and-rise-again-of-the-first-ai-agent-millionaire/

'Can we really trust AI to channel the public's voice to ministers?' *The Guardian*, https://www.theguardian.com/commentisfree/2024/apr/25/ai-public-voice-ministers-large-language-model-chatgpt

'Seth Lazar on Legitimate Power, Moral Nuance, and the Political Philosophy of AI', *Generally Intelligent*, podcast interview, https://open.spotify.com/episode/0xRSqLMLLqR1z6jyK5THBU?si=vLKC1jUdTsiG_eTNWQQMoA

2023 'The US is racing ahead in its bid to control artificial intelligence – why is the EU so far behind?', *The Guardian*, https://www.theguardian.com/commentisfree/2023/nov/28/united-states-artificial-intelligence-eu-ai-washington

'Model alignment protects against accidental harms, not intentional ones', with Arvind Narayanan and Sayash Kapoor, *AI Snake Oil*, https://www.aisnakeoil.com/p/model-alignment-protects-against

'Political Philosophy in the Age of AI', *Philosophy Bites* Interview, https://philosophybites.libsyn.com/seth-lazar-on-political-philosophy-in-the-age-of-ai

'AI: Is it Out of Control? , *Science Vs* Interview, https://open.spotify.com/episode/3NV59n4abm9JBV0sWkEDyi?si=s8hA8oHiS8ivmHHPr7_dqw

'Is Avoiding Extinction from AI Really an Urgent Priority?' with Arvind Narayanan and Jeremy Howard, *AI Snake Oil*, https://www.aisnakeoil.com/p/is-avoiding-extinction-from-ai-really

'Machines and Morality', New York Times, https://www.nytimes.com/2023/06/19/special-series/chatgpt-and-morality.html

2020 'Large-Scale Facial Recognition is Incompatible with a Free Society', with Claire Benn and Mario Günther, *The Conversation*, https://theconversation.com/large-scale-facial-recognition-is-incompatible-with-a-free-society-126282

2020 'Contact tracing apps are vital tools in the fight against coronavirus. But who decides how they work?', with Meru Sheel, *The Conversation*, https://theconversation.com/contact-tracing-apps-are-vital-tools-in-the-fight-against-coronavirus-but-who-decides-how-they-work-138206

2020 'We're Sleepwalking into a World of Mass Surveillance', *Barron's*, https://www.barrons.com/articles/your-consent-to-contact-tracing-apps-is-meaningless-51588250580

2019 AI and moral intuition: use it or lose it? ABC Philosopher's Zone interview, https://www.abc.net.au/listen/programs/philosopherszone/ai-and-moral-intuition:-use-it-or-lose-it/12075060

2019 Interviewed on Episodes 2 and 4 of Series 3 of HiPhi Nation, approx. 100,000 downloads

2018 'Why we need more than just data to create ethical driverless cars' (with Colin Klein), *The Conversation*, http://theconversation.com/why-we-need-more-than-just-data-to-create-ethical-driverless-cars-105650

2016 'Should Civilians be Spared?' *Examining Ethics*, podcast, https://www.prindleinstitute.org/podcast/should-civilians-be-spared/

2014 On Human Shields, *Boston Review*, https://www.bostonreview.net/articles/seth-lazar-human-shields/

2014 'Sparing Civilians in War' *Philosophy Bites*, interview by Nigel Warburton and David Edmonds. Released 19/7/2014, https://philosophybites.libsyn.com/seth-lazar-on-sparing-civilians-in-war

2013 The Moral Responsibility of Volunteer Soldiers: Response to McMahan, *Boston Review*, https://www.bostonreview.net/forum/moral-wounds-ethics-volunteer-military-service/

2012 'Seth Lazar on Self-Defense in War' *Public Ethics Radio*, Christian Barry (ed.), Carnegie Council for Ethics in International Affairs, https://www.carnegiecouncil.org/media/series/004/20120316-seth-lazar-on-self-defense-in-war

Grants

External Grants

2024 A Conceptual and Practical Toolkit for Sociotechnical AI Safety, Center for Security and Emerging Technology, USD500,000
Sole CI. Two-year project, deferred to start when I arrive at Johns Hopkins.

2024 Language Model Agents and Society Project, Templeton World Charity Foundation, USD999,886
Sole CI. Three-year project on anticipating and steering the societal impacts of Language Model Agents.

2024 Support for Machine Intelligence and Normative Theory Lab, Survival and Flourishing DAF, USD 480,000
Sole CI. Unrestricted funding to support work on sociotechnical AI safety.

2024 'Implementing a "Moral Conscience" for LLM Agents', OpenAI Agents Research Grant — USD50,000
Lead CI, with Dylan Hadfield-Menell, Daniel Kilov and Aaron Snoswell.

2023 'Normative Philosophy of Computing' field building grant. AI2050 — USD50,000
Sole CI. Grant to build up the field of normative philosophy of computing.

2022 'Social Ontology of Large Language Models'. Google Research Unrestricted Gift — USD10,000
Small grant to support work on societal impacts of LLMs.

2022 'Socially Responsible Insurance in the Age of AI'. ARC Linkage Projects (LP21) Award — AUD495,000 with AUD350,000 funding from IAG and AUD100,000 from ANU
Lead CI, with Damian Clifford, Jenny Davis, Kimberlee Weatherall, Tiberio Caetano and Chris Dolman. A collaboration with the Gradient Institute and Insurance Australia Group to establish how AI can be used to help realise the social function of insurance while mitigating risks due to discrimination, unaccountable power, and privacy.

2021 'Automatic Authorities: Charting a Course for Legitimate AI'. ARC Future Fellowship (FT21) Award — AUD1,020,698
Sole CI. Project commenced April 2022. This project aims to develop novel theories of power and its justification, and apply them to the use of AI by state and non-state actors to exercise power by means of AI.a

2019 'Moral Skill and Artificial Intelligence'. Templeton World Charity Foundation 'Diverse Intelligences' Grant — USD234,000
Project Director. With Co-Director Claire Benn, and CIs Jenny Davis, Toni Erskine, Colin Klein. This project will ask whether outsourcing morally weighted decisions to automated systems can lead to 'moral deskilling', and whether there are ways to design automated systems so that they make us better, not worse, at exercising moral judgment.

2016 'Ethics and Risk'. ARC Discovery Project Award (DP17) — AUD335,000
Lead CI. Other CIs and PIs: Lara Buchak (Berkeley); Katie Steele (ANU); Alan Hájek (ANU); Frank Jackson (ANU); Philip Pettit (ANU).

2013 'Justifying War'. ARC Discovery Early Career Research Award (DE13) — AUD366,000
Sole CI.

05-08 Arts and Humanities Research Council, Doctoral Fellowship — GBP60,000

Major Internal Grants

2019 'Humanising Machine Intelligence'. ANU Grand Challenge Program (AUD5.5m)
Founding Lead (stepped down to take up future fellowship). Other team members: Damian Clifford, Jenny Davis, Toni Erskine, Colin Klein, Hanna Kurniawati, Sarah Logan, Katie Steele, Sylvie Thiébaux, Lexing Xie. Multidisciplinary project on the morality, law and politics of data and AI. For information see hmi.anu.edu.au

Research Events Convened

Conferences

2022 ACM Fairness, Accountability and Transparency Conference
General co-chair (one of four) for the top CS and interdisciplinary conference for AI ethics.

2021 AAAI/ACM AI, Ethics and Society Conference
Program and General co-chair (one of four) for one of the two top CS and interdisciplinary conferences for AI ethics. Also convened 'Platform Power and AI' panel discussion.

Workshops Organised

2025 Sociotechnical AI Safety Retreat at Kioloa Coastal Campus; Knight 1A Symposium on AI and Democratic Freedoms.

2024 Normative Philosophy of Computing Workshop at Kioloa Coastal Campus; AI and Catastrophic Risk at ANU; Sociotechnical AI Safety Workshop at ITS Rio; Political Philosophy and AI at Kioloa; Normative Philosophy of Computing at Yale; Knight 1A Symposium on AI and Democratic Freedoms.

2023 Philosophy, AI and Society Workshop and Fair Machine Learning Authors meet Critics Workshop at Stanford; Philosophy AI and Society Doctoral Colloquium at Oxford; Sociotechnical AI Safety Workshop at Stanford; Democracy and AI Workshop at Carnegie Endowment for International Peace.

2022 Philosophy, AI and Society workshop at Harvard

2019 'Ethics and AI'; 'Decision Theory and AI'; FM Kamm Masterclass; Stanford HAI Philosophy, AI, and Society panel and workshop at Stanford

2018 'Ethics and Risk'; 'Foundations of Normative Ethics'

2017 'On the work of Marc Fleurbaey'; 'Awesome Workshop in Normative Ethics'

2016 'Awesome Workshop in Normative Ethics and Political Philosophy'; PPE Masterclass; Dale Dorsey Masterclass; Jake Ross Masterclass; Wlodek Rabinowicz Masterclass

2015 'Ethics and Decision Theory'

2014 'Feasibility and the Ethics of War' (with Nic Southwood); AAP Stream on Ethics of Force; Christian List Masterclass

2013 Honours/Masters Workshop; Legitimacy and Authority; Niko Kolodny Masterclass; Enoch/Jackson/Smith Masterclass; Tom Dougherty Masterclass

2011 'War and Global Justice', IAS, Hebrew University of Jerusalem

2010 'Why We Fight: The Purposes of Military Force in the Twenty-First Century', second meeting of the ELAC Workshop, Oxford; 'Eliminative and Manipulative Agency in the Ethics of Self-Defence', ELAC, Oxford

2009 '*Killing in War* workshop', first meeting of the ELAC Workshop, Oxford

Public Lectures Organised

2022 Jamie Susskind launching the Machine Intelligence and Normative Theory (MINT) Lab, https://philosophy.cass.anu.edu.au/index.php/news/jamie-susskind-lecture

2019 Jack Smart lecture by FM Kamm; HMI Lectures by Walter Sinnott-Armstrong, Shannon Vallor, David Danks, Kate Crawford

2018 Jack Smart lecture by Michael Smith; Philosophy and Public Policy Lecture by Huw Price

2017 Jack Smart lecture by Peter Godfrey-Smith; Launch of Centre for Philosophy of the Sciences, public lecture by Peter Godfrey-Smith; Passmore Lecture by Marc Fleurbaey; Philosophy and Public Policy Lecture by Peter Singer

2016 Philosophy and Public Policy Lecture by Leif Wenar; Passmore Lecture by Liz Anderson

2014 Passmore Lecture by Jeff McMahan

Invited Presentations

Named Lectures

'On AI Personhood Without Sentience', Arthur and Barbara Gianelli Annual Lecture, **St John's University**, 2025

'What, if anything, should we do, now, about catastrophic AI risk?'. Scholl Lecture, **Purdue University**, 2024.

'Algorithmic Governance and Political Philosophy'. Tanner Lectures on AI and Human Values at the Human-Centered AI Institute, **Stanford University**, 2023. Recordings.

'The Nature and Justification of Algorithmic Power'. Mala and Solomon Kamm Lecture in Ethics at the Safra Centre for Ethics, **Harvard University**, 2022. Recording.

Keynotes and Festschrift

'Themes from Seth Lazar'. Half-day workshop on my work at **University of Hong Kong**, 2025.

'Evaluating LLM Ethical Competence'. **NeurIPS Workshop on Algorithmic Fairness through the Lens of Metrics and Evaluation**, Vancouver, 2024. Recording.

'Philosophical Foundations for Pluralistic Alignment'. **NeurIPS Workshop on Pluralistic Alignment**, Vancouver 2024

'What, if anything, should we do, now, about catastrophic AI risk?'. **European Workshop on Algorithmic Fairness**, Mainz, 2024.

'Legitimacy, Authority, and the Political Value of Explanations'. **Oxford Studies in Political Philosophy Conference**, Arizona, 2022.

'Legitimacy, Authority, and the Political Value of Explanations'. Japan Association for Philosophy of Science Annual Meeting, **Tokyo Institute of Technology**, 2021.

'On Machine Ethics'. Keynote lecture at Ethics of Data Science Conference, **University of Sydney**, 2019

Keynote Lecture at Ethics of War in the 21st Century, **University of Stockholm** (2014).

Departmental Seminars and other Invited Talks

'Aligning Language Model Agents', Lingnan University AI Ethics Workshop.

'What, if anything, should we do, now, about catastrophic AI risk?' 2024: Hong Kong University, Sociotechnical AI Safety Workshop in Rio.

Aligning LLM Agents', 2024: OpenAI, Purdue University, Stanford School of Engineering.

'Governing the Algorithmic City' (previously, 'The Nature and Justification of Algorithmic Power'). 2023: Cal Poly; 2022: Cornell Tech Digital Life Institute, Princeton University Center for Human Values, Carnegie Mellon, Emory, Toronto Schwarz Reisman Institute for Technology.

'Communicative Justice and the Distribution of Attention'. Cornell, Notre Dame, Knight First Amendment Institute Symposium on Algorithmic Amplification, Columbia.

'Legitimacy, Authority, and the Political Value of Explanations'. 2022: Rutgers University; 2020: Diverse Intelligences Summit; 2019: ANU, MIT, Carnegie Mellon.

'On the Site of Predictive Justice'. 2023: University of Oxford Institute for Ethics in AI; 2022: Princeton Workshop in Normative Philosophy, Edmond J. Safra Center for Ethics, Harvard, ANU.

'What's Wrong with Automated Influence'. 2020: Edmond J. Safra Center for Ethics, Harvard.

'Machine Ethics: A Solution in Search of a Problem'. 2020: Carnegie Mellon; 2018: ANU.

'AI Ethics without Principles'. 2020: Brown Bag Lecture Series, World Bank, Washington DC (Cancelled due to COVID); 2019: NITRD Agency

Previous talks: Moral Philosophy Seminar, **University of Oxford** (2017, 2013); Moral Sciences Club, **University of Cambridge** (2017); **National University of Singapore** (2017); **University of York** (2017); **University of Melbourne** (2016, 2011); **MIT** (2015, 2019); **UNC-Chapel Hill** PPE Group (2015); **Yale** Moral Philosophy Working Group (2015, 2013); Kadish Center for Morality, Law and Public Affairs, **UC Berkeley** (2015); **Monash University** (2015); **Arizona State University** (2015); **University of Southern California** (2015, 2013); **UNC-Greensboro** (2015); Nathanson Centre for Human Rights, **York University**, Toronto (2014); CSMN Research Seminar, **University of Oslo** (2014); **University of Toronto** (2014, 2011); Program in Ethics and Public Affairs Seminar, **Princeton** (2014); Roskilde University, Copenhagen (2014); **University of St Andrews** (2013); **Victoria University, Wellington** (2013); Uehiro Centre for Practical Ethics, **University of Oxford** (2013); **University of Stirling** (2013); **University of Auckland** (2013); **University of Otago** (2013); **University of Manchester** (2013); **University of Glasgow** (2013); **University of Christchurch** (2013); **Macquarie University** (2013); **University of Adelaide** (2013); **Stanford University** (2013, 2011); **University of Essex** (2013); **University of Warwick** (2013, 2010); **Rutgers University** (2013); **University of Leeds** (2013); **University of Colorado at Boulder** Political Science Department (2013); Political Science, **LSE** (2013); Political Theory Seminar, **University of Cambridge** (2013); Political Theory Seminar, **Cambridge** (2013); **Dartmouth Department of Government** (2011); **Washington University St Louis** (2011); **University of Chicago** (2011); **University of Toronto** (2011); Nuffield Political Theory Workshop, **University of Oxford** (2011).

Conferences/Workshops

ACM FAccT Tutorial 2024; **ACM FAccT** Tutorial, 2021; Modelling Morality Workshop, **Carnegie Mellon University** (2020)**;** Ethics and Uncertainty Conference, Center for Moral and Political Philosophy, **Hebrew University of Jerusalem** (2018); Ethics and Risk Workshop, **ANU** (2018); **Social Philosophy and Policy** workshop, London (2018); **Oxford Studies in Normative Ethics** Workshop, Tucson (2018); New Work in Political Philosophy workshop, **Hebrew University of Jerusalem** IAS (2016); Workshop on Ethics and Risk, **University of Stockholm** (2016); ICREA Public Health Workshop, **Barcelona Pompeu-Fabra** (2016); US Military Academy, **West Point** (2015); American Academy Conference, New Dilemmas in Ethics, Technology and War I, **Stanford University** (2015); American Academy Conference, New Dilemmas in Ethics, Technology and War II, USMA, **West Point** (2015); Oxford Studies in Political Philosophy Workshop, **Syracuse** (2015); Feasibility and the Ethics of War Workshop, **ANU** (2014); CRNAP Oxford-ANU-Princeton Workshop, **ANU** (2014); Ethics and War Conference, **UCSD** (2013); Workshop at **Stockholm Centre for the Ethics of War and Peace** (2014); Self-defence workshop, Centre for Human Values, **Princeton** (2013); 'Ethics and Law in War', ELAC workshop, **University of Oxford** (2011); 'Why We Fight: The Purposes of Military Force', ELAC workshop, **University of Oxford** (2010); Workshop on Ethics, Jus Post Bellum, and International Law, CAPPE, **ANU** (2010); 'Asymmetric Wars, International Relations, and Just War Theory', **Belgrade University** (2010); Oxford and Princeton Global Norms/Global Justice Research Collaboration, **University of Oxford** (2009).

Research Impact

2025 Platform Agents research informed policy consultation with industry lobbyists, senate staffers, and Consumer Reports, Power and AI research extensively engaged with in UNDP Human Development Report,

https://hdr.undp.org/system/files/documents/global-report-document/hdr2025reporten.pdf

2023 Contributing Author to Rapid Response Information Report on Generative AI by Australian Council of Learned Academies

21-22 Member of 14-person US National Academies of Science, Engineering and Medicine Study Committee on 'Ethics and Governance of Responsible Computing Research', resulting in publication of 150-page co-authored report, led by Professor Barbara Grosz (Harvard)

2021 Contributing Author to Rapid Research Information Forum on Motivators for use of the COVIDSafe App

2020 Public Submission (with other members of HMI and Australian Academy of Science) to Human Rights Commission in response to Human Rights and Technology Report

2020 Participated in Gradient Institute-led workshop on the ethics of insurance pricing with IAG

2019 Public Submission (with Bob Williamson and the Australian Academy of Science) to the DIIS consultation on Australia's Ethical Framework for AI.

2019 Invited Speaker at Plan Jericho/Trusted Autonomous Systems/Defence Science and Technology workshop on Ethical AI for Defence

2019 Invited Participant at Defense Innovation Board Roundtable on AI policy principles for the US Department of Defense, subsequently provided comments on draft principles

2019 Invited written submission for Defense Innovation Board consultation on AI policy principles

2019 Invited Participant, Defence Science and Technology Megatrends Workshop

2019 Expert Working Group member, Academy of Social Sciences of Australia report to Office of National Intelligence on US National Academies Decadal Survey

2019 Public Submission (with Bob Williamson and the Australian Academy of Science) to the Data61/DIIS consultation on Australia's Ethical Framework for AI

2018 Invited Public Submission to Defense Innovation Board consultation on AI policy principles for the US Department of Defense

2015 Invited participant in MacArthur Foundation and American Academy of Arts and Sciences project on New Technologies and the Ethics of War

2010 Contributed to Advisory Consultation, US Army Professional Military Ethics code

HDR Supervision

2011- PhD Panel Chair, School of Philosophy, ANU
Kira Breithaupt (started 2025), Iman Ferestade (started 2025), Tim Dubber (started 2025), Andrew Smart (started 2024), Jake Stone (started 2021), Max Fedoseev (PhD 2024), Josef Holden (PhD 2023), Chad Lee-Stronach (PhD 2019), Adam Gastineau (MPhil 2016). Fedoseev and Lee-Stronach both won VC Teaching Awards for tutoring. Lee-Stronach is now TT at Northeastern after a Stanford Postdoc. Stone has been offered a postdoc working with Sandra Wachter.

2011- PhD Panel Member, School of Philosophy, ANU
Emily Leijer, James Lim, Chris Lernpass, Jenny Munt, Shalom Chalson (PhD 2025), Kirsten Mann (PhD 2024), Devon Cass (PhD 2020), Heather Browning (PhD 2020), Ten Herng Lai (PhD 2020), Adam Bugeja (PhD 2018), Rob Kirby (PhD 2017), Lachlan Umbers (PhD 2017), Matt Hammerton (PhD 2016), Chris Gyngell (PhD 2015), Jonathan Pickering (PhD 2014), Jo Lau (PhD 2013), Stephanie Collins (PhD 2012).

2011- PhD Panel Member, Other
Member of dissertation committee for Adam Betz (University of Illinois, Chicago, PhD 2016)*, and Steve Woodside (Rutgers, PhD 2016)*.

Coursework Teaching

2015- Honours Supervision
Aleks Hammo (2023 H1), Antonio Esposito (2023), Matthew Wiseman (2020 H1), Kida Lin (2019 H1), Eleanor Kay (2018 H1), Oliver Rawle (2017 H1), Kramer Thompson (2017 H1), Julian Christopher Scott (2015 H1).

2019 Co-Convenor of PHIL3073 'The Moral and Political Philosophy of AI'. 70 students.

2019 Foundations Graduate Seminar (PHIL8011) on 'The Philosophy of AI'. 16 students.

2019 Co-Convenor of ANU/Humboldt/Princeton Summer Institute on Normativity. 23 students.

2017 Foundations Graduate Seminar (PHIL8011) on 'Moral Decision Theory'. 16 students.

15-16 Convenor of Philosophy Honours.

13-15 Convenor of PHIL8011 Foundations Seminar.

2013 Foundations Graduate Seminar (PHIL8011) on 'Liberty', with Philip Pettit. ≈16 students

11-14 Convened Graduate Work-in-Progress group in MSPT.

2014 Introduced and organised new Project Design Review for first year MSPT students

07-09 Pembroke College, Oxford
Designed syllabus and taught tutorials, for courses on Kant's Ethics, War and Global Justice, Marx and Marxism, and Theory of Politics and Ethics. Also taught revision seminars (larger classes). One of my students went on to become an academic philosopher, now at Cardiff.

06-09 Regent's Park, St. Hugh's, St. Peter's, St. Hilda's, Wycliffe Hall, Oxford
Designed syllabus and taught tutorials for War and Global Justice, Theory of Politics, Ethics

06-09 Oxford Overseas Study Course, Taylor University Programme, Oxford

Service (Profession)

25- Section Editorial Committee member for Computing Research Repository (arXiv)

25- Associate Editor, *Philosophy & Public Affairs*

22-25 Executive Committee of ACM Fairness, Accountability and Transparency Conference

24-25 Expert recommender and speculation grantor focused on AI Safety for Survival and Flourishing Fund, reviewer for AI Safety Fund and UK AI Safety Institute.

21- Area Chair: NeurIPS (2025), ICML (2025), CoLM (2024, 2025), FAccT (2022, 2023, 2024, 2025)

21- Program Committee member: AIES (2022, 2023), IJCAI (2021)

21-23 Devised and directed Philosophy, AI and Society Consortium of associated universities, uniting philosophers at ANU, Oxford, Stanford, Toronto, Princeton and Harvard.

19-23 Presidential Nominee on MIT Corporation Visiting Committee for the Department of Linguistics and Philosophy

2019 Expert Working Group, Academy of Social Sciences of Australia, report to Office of National Intelligence on *Social Science Research & Intelligence in Australia*

18-25 Editor, *Philosophers' Imprint*

2016- Editorial Board, *Oxford Studies in Political Philosophy*

15-19 Area Editor (political philosophy), *Ergo*

2012- Editorial Board, *Journal of Political Philosophy*

13-14 Social and Political Philosophy editor for PhilPapers

2010 Oxford University Press, Oxford Bibliographies Online
Contributed 'War' Entry to online bibliography

2008- Referee: Ethics, Law and Philosophy, Social Theory and Practice, British Journal of Political Science, Political Studies, International Theory, Journal of Moral Philosophy, Philosophy and Psychology, Philosophical Studies, Review of International Studies, Politics, Philosophy and Economics, Journal of Applied Philosophy, Res Publica, Canadian Journal of Philosophy, Ratio, European Journal of Political Theory, Philosophical Quarterly, Philosophical Review,

2008- Refereeing and commentary: Oxford University Press, Routledge, Chicago University Press