Patrick Butlin

Global Priorities Institute, University of Oxford
patrick.butlin@gmail.com +44 7906210518

AOS: Philosophy of Cognitive Science (AI, Psychology & Neuroscience), Philosophy of Mind
AOC: Ethics, Logic

Employment

2024- *Global Priorities Institute, University of Oxford*
Postdoctoral Research Fellow

2021-2024 *Future of Humanity Institute, University of Oxford*
Research Fellow

2017-2021 *King's College London*
Visiting Research Fellow in Philosophy (2020-21)
Teaching Fellow in Philosophy (2017-20)

2016-2017 *Centre for Philosophical Psychology, University of Antwerp*
Postdoctoral Fellow

2014-2017 *Hertford College, University of Oxford*
Stipendiary Lecturer in Philosophy

Education

2011-2015 *King's College London*
PhD Philosophy (Supervisors: Prof. David Papineau, Prof. Nicholas Shea)

2010-2011 *University of Sheffield*
MA Cognitive Studies – Distinction

2008-2010 *Merton College, University of Oxford*
BPhil Philosophy

2003-2007 *Merton College, University of Oxford*
MMathPhil Mathematics and Philosophy – First Class

Publications

AI Consciousness and Welfare Reports:

'Taking AI Welfare Seriously'. *arXiv:2411.00986*. 2024.

[10 authors. Media coverage included *The Times*, *The Guardian* and BBC Radio 4.]

'Consciousness in Artificial Intelligence: Insights from the Science of Consciousness'. *arXiv:2308.08708*. 2023.

[30,000-word report; joint first author of 19 including neuroscientists, AI researchers and philosophers. Media coverage included news articles in *Science*, *Nature* and *New Scientist*.]

Full-Length Journal Articles:

'Principles for Responsible AI Consciousness Research' *Journal of Artificial Intelligence Research*, forthcoming.

'The Agency in Language Agents' *Inquiry* (SI: Mind, Language and AI), forthcoming.

'AI Assertion' (with Emanuel Viebahn). *Ergo*, forthcoming.

'Reinforcement Learning and Artificial Agency'. *Mind & Language* 39(1): 22-38, 2024.

'Machine Learning, Functions and Goals'. *Croatian Journal of Philosophy* (papers from Kathy Wilkes Memorial Conference) 22(3): 351-370, 2022.

'Sharing Our Concepts with Machines'. *Erkenntnis* 88: 3079-3095, 2023 (online 2021).

'Directive Content'. *Pacific Philosophical Quarterly* 102(1): 2-26, 2021.

'Cognitive Models are Distinguished by Content, not Format'. *Philosophy of Science* 88(1): 83-102, 2021.

'Affective Experience and Evidence for Animal Consciousness'. *Philosophical Topics* 48(1): 109-127, 2020.

'Representation and the Active Consumer'. *Synthese* 197: 4533-4550, 2020 (online 2018).

'Why Hunger is not a Desire'. *Review of Philosophy and Psychology* 8(3): 617-635, 2017.

Comments, Chapters and Conference Proceedings:

'Can reinforcement learning model learning across development? Online lifelong learning through adaptive intrinsic motivation' (with Kai Sandbrink, Brian Christian, Linas Nasvytis and Christian Schroeder de Witt). *Proceedings of the Annual Meeting of the Cognitive Science Society* 46, 2024.

'Sentience Criteria to Persuade the Reasonable Sceptic' (comment on Crump et al., 'Sentience in Decapod Crustaceans'). *Animal Sentience* 32(21), 2022.

'AI Alignment and Human Reward'. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics and Society (AIES '21)* 437-445, 2021.

'Normal and Addictive Desires' (with David Papineau). In Nick Heather and Gabriel Segal (eds.), *Addiction and Choice: Rethinking the Relationship*. OUP, 2016.

Presentations

Refereed Presentations:

'Agency and Imitation'
- Philosophy of Science Association, New Orleans, 15 November 2024

'Applying the Science of Consciousness to Current AI Systems'
- Association for the Scientific Study of Consciousness, NYU, 23 June 2023

'Functions, Content and Understanding in Large Language Models'
- Philosophy of Deep Learning Conference, NYU, 26 March 2023

'Valence and Reflexivity'
- 3rd Joint Meeting of ESPP and SPP, Milan, 21 July 2022

'Well-Being, Agency and the Individual Evaluative Perspective'

- British Society for Ethical Theory Annual Conference, 13 July 2022

'Agency, Learning, Functions and Goals'

- Workshop on RL as a Model of Agency, RLDM, 11 June 2022

'Model-Based Reinforcement Learning and Action for Reasons'

- Cambridge CFI Kinds of Intelligence Workshop, 21 January 2022

'AI Alignment and the Complex Structure of Human Values'

- CEPE/IACAP Joint Conference: The Philosophy and Ethics of AI, 7 July 2021

'Reward, Reward Signals, and Agents within Agents'

- 47th Annual Meeting of the Society for Philosophy and Psychology, 30 June 2021

'AI Alignment and Human Reward' (poster)

- AIES '21, 19-21 May 2021

'Sharing Our Conceptual Resources with Machines'

- PLM5, St. Andrews, 29 August 2019
- Kinds of Intelligence 2: Machine Minds, Cambridge, 20 June 2019

'Model-Based Reinforcement Learning and Acting for Reasons'

- Anticipatory Systems: Humans meet AI, Örebro, 10 June 2019

'Pleasure, Pain and the Evolution of Consciousness'

- European Society for Philosophy and Psychology, Rijeka, 13 September 2018

'Information-how and Directive Content'

- Joint Session, Oxford, 8 July 2018

'Liberal Representation and the Active Consumer'

- ECMN Future Minds Conference, Warwick, 16 March 2017

'Liberal Representation and Basic Action'

- 10th Logos Workshop on Naturalistic Theories of Intentionality, Barcelona, 1 December 2016

'Lewis and Anscombe on Direction of Fit'

- Early Career Mind Network, Durham, 1 September 2016

'Naturalizing Direction of Fit'

- KCL-UNC Workshop on 'The Normative in a Natural World', UNC-Chapel Hill, 3-5 May 2013

'Teleosemantics and the Direction of Fit of Desire'

- Graduate Conference in Theoretical Philosophy, Groningen, 18-20 April 2013

'Holton's Argument from Strength of Will'

- European Society for Philosophy and Psychology, London, 28-31 August 2012

'What's in a That-Clause?' (with Emanuel Viebahn)

- BPPA Conference, University of Edinburgh, 3-6 September 2012

'Schroeder's *Being For* and the Structure of the Propositional Attitudes'

- Oxford Philosophy Graduate Conference, 20-21 November 2010

Invited Presentations:

'AI Assertion in 2025'

- Uppsala Vienna AI Colloquium, 17 January 2025

'Could Language Agents have Desires?'

- LAIHP AI and Affect Seminars, London, 22 October 2024

'AI Sentience: New Directions'

- Effective Altruism Global, London, 1 June 2024

'Reflections on Consciousness in AI'

- Turing Fringe Lectures, Queen's University Belfast, 16 April 2024

'Could a 'Language Agent' Really be an Agent?'

- 2nd Recife Virtual Conference on Philosophy of Mind, 20 February 2024

'Consciousness in AI: Insights from the Science of Consciousness'

- Google DeepMind, London, 24 October 2023
- Hong Kong University, 11 October 2023
- Ezra Hale Lectures in AI, Rochester Institute of Technology, 3 October 2023
- Umeå University, 29 September 2023
- AI Safety Hub Edinburgh, 14 September 2023

'Functions, Content and Understanding in Language Models'

- Workshop on Philosophy of Large Language Models, Eindhoven, 26 January 2023

'Individual Evaluative Perspectives'

- Rules of Attraction Workshop, Tartu, 16 December 2022

'Artificial Agency'

- Institute for Ethics in AI Seminar, Oxford, 23 November 2022

'AI Safety and Artificial Agency'

- UCL Effective Altruism Society, 2 March 2023
- Oxford University Philosophy and AI Societies, 14 November 2022

'Thoughts on AI, Art and Agency'

- AI Art Salon, National Gallery X, 25 May 2022

'Speech Acts in AI' (with Emanuel Viebahn)

- Jülich Research Group, 13 May 2022

'Machine Learning, Functions and Goals'

- Kathy Wilkes Memorial Conference, IUC Dubrovnik, 29 April 2022

'The Palermo Protocol as Evidence for Consciousness'

- Workshop on Non-Human Consciousness, CFI Cambridge, 21 April 2022

'Addiction, Hijacking and Autonomy'

- KCL Neuroscience Society, 26 November 2019

'Explaining Consciousness'

- Consciousness: Neuroscience v. Philosophy, KCL Neuroscience Society, 21 March 2019

'Imperative Intensity'

- Direction of Fit Workshop, Antwerp, 9 November 2016

'Desire as a Natural Kind'

- A Highly Desirable Symposium, Durban, 18 June 2016

'The Direction of Fit of Desire'

- Hertford Philosophical Society, Oxford, 18 February 2016

'Why Hunger is Not a Desire'

- London-Warwick Mind Forum, UCL, 16 June 2015

'Normal and Addictive Desires' (with David Papineau)

- KCL-UNC Workshop on 'Philosophical Approaches to Desire and Pleasure', King's College, London, 19-20 May 2014

'What's in a That-Clause?' (with Emanuel Viebahn)

- Ockham Society, University of Oxford, 12 June 2012

Comments:

On Ethan Jerzak, 'Non-Classical Knowledge'

- London-Berkeley Graduate Conference, UC-Berkeley, 10-11 May 2013

On Sam Wilkinson, 'Explaining Addiction'

- KCL Graduate Conference in Philosophy of Mind and Psychology, 26 April 2013

Teaching

As a Teaching Fellow at King's College London:

Module Convenor and Lecturer:

- Advanced Topics in Philosophy of Mind (Neuroscience and Biomedical Science students; Spring 2019 & 2020)
- Neuroscience and the Mind (philosophy of mind for Neuroscience and Biomedical Science students; Autumn 2018 & 2019)
- Topics in Philosophy of Psychology (BSc Psychology students; Spring 2018)
- Philosophy of Psychology (BSc Psychology students; Spring 2018)
- Philosophy of Psychology (BA and MA Philosophy students; Autumn 2017)

Other classes:

- Neuroscience and the Mind (seminars and revision lectures 2017-18)
- Philosophy for medical students (seminars)
- Ethics I (seminars)

Dissertation supervision for 7 undergraduates, 6 MA students

Tutorials and Classes at Oxford, 2014-2017 and 2022-2024:

- Philosophy of Mind (5 students)
- Philosophy of Cognitive Science (7 students)
- Ethics (18 students, 3 visiting students)
- Elements of Deductive Logic (12 students)
- Moral Philosophy (first year) (18 students)

Supervision for one BPhil student

Impact and Public Engagement

Invited participant, GESDA (Geneva Science and Diplomacy Anticipator) Future of Consciousness Anticipation Workshop, 20 June 2024

Participant, 'AI Consciousness Report: A Roundtable Discussion', NYU Mind, Ethics and Policy Program, 5 September 2023

Discussions with Miele and Z_punkt staff on future of digital technology, 14 September 2022

Participant, AI Art Salons, National Gallery X, 10 May and 25 May 2022

Judge, King's College London Regional, John Stuart Mill Cup 2021 & Ethics Cup 2022

Participant, panel on 'Addiction: What, Who and How?' at *Hooked* exhibition, Science Gallery London, 19 October 2018

Research Grant

Jan-Jun 2021 *Survival and Flourishing*
Support for research on AI alignment and human values ($19,200)

Awards and Funding

AI Consciousness Project:

EA Long-Term Future Fund Grant for Montreal Workshop ($10,840, April 2023)
Support from Effective Ventures for Oxford workshop (c. £5000, December 2022)

University of Oxford:

OXBER Grant for Oxford-Berlin research collaboration, with Emanuel Viebahn (€1200, Autumn 2021)

King's College, London:

King's Undergraduate Research Fellowship (funding for undergraduate research assistant, Summer 2019)
AHRC Doctoral Award (Fees & Full Maintenance, 2011-2014)
AHRC Research Training and Support Grant for visit to Cambridge University (£1092.40, Autumn 2013)

University of Edinburgh:

Principal's Career Development Scholarship (£14000-15000 p.a., 2011-2014) – *declined*

University of North Carolina, Chapel Hill:

Richard Brooke Fellowship (Non-teaching stipend for PhD from 2008) – *declined*

Merton College, University of Oxford:

Postmastership 2005-2007 (highest academic scholarship)
Exhibition 2004-2005 (academic scholarship)

Professional and Departmental Service

Refereeing for publication:

Books: *Oxford University Press; SAGE Publishing*
Journal articles: *AI & Society; American Philosophical Quarterly; Analysis; Australasian Journal of Philosophy; British Journal for the Philosophy of Science; Biology & Philosophy; Episteme; Erkenntnis; European Journal of Analytic Philosophy; Inquiry; Journal of the American Philosophical Association; Journal of Consciousness Studies; Mind; Mind & Language; Neuroscience of Consciousness; Pacific Philosophical Quarterly; Philosophical Psychology; The Philosophical Quarterly; Philosophical Studies; Philosophy; Philosophy and the Mind Sciences; Philosophy and Phenomenological Research; Philosophy of Science; Physics of Life Reviews; PNAS; Review of Philosophy & Psychology; Synthese; Theoria*

Departmental service at King's College, London:

- Study abroad tutor

- External relations lead
- Convenor, PGR and MA research seminars
- Personal tutor and liaison tutor for Modern Languages and War Studies
- Web and social media lead

Mentor, Summer Research Fellowship Programme, Principles of Intelligent Behavior in Biological and Social Systems, 2022-2024

Tutor, University of London Philosophy of AI Summer School, 2024

Admissions interviewer, Hertford College, 2014 & 2015

Conference organisation:

- AI Consciousness Project Workshops, Oxford/online, 2 December 2022 and Montreal/online, 27 April 2023
- FHI/KCL Workshop on the Grounds of Moral Status, online, 20-21 June 2022
- FHI Workshop on Desire, Valence and Affect in AI, online, 22-23 November 2021
- Direction of Fit Workshop, Antwerp, 9 November 2016
- 5th Annual UNC-KCL Workshop on 'Philosophical Approaches to Desire and Pleasure', 19-20 May 2014

Advisory Position and Non-Academic Employment

2018-2020 Advisor on philosophy and cognitive science at Sphere Knowledge
2007-2008 Research Analyst at ECA International