# Steve Petersen

Professor of Philosophy
Niagara University
Department of Philosophy
PO Box 2043
Niagara University NY 14109

716-406-7763
steve|at|stevepetersen.net
https://stevepetersen.net
last updated: November 13, 2025

## AREAS

### Areas of specialization

Philosophy and ethics of artificial intelligence, philosophy of mind, philosophy of science

### Areas of competence

Epistemology, logic, metaphysics, philosophy of language

## EDUCATION

### University of Michigan

PhD, philosophy

*Belief-Desire Coherence*
Committee: Eric Lormand (chair), Jessica Wilson, James Joyce, Thad Polk.

### Harvard University

BA, philosophy with mathematics

## PUBLICATIONS

"Abstractions by patterns" (forthcoming), *Real Patterns in Science and Nature*, ed. Tyler Millhouse, Steve Petersen, and Don Ross. Cambridge: MIT Press.

"Multiple patterns, multiple explanations" (2023), *Conjunctive Explanations: New Essays on the Nature, Epistemology, and Psychology of Explanatory Multiplicity*, ed. Jonah Schupbach and David Glass. New York: Routledge.

"In defence of the hivemind society" (2021), with John Danaher, *Neuroethics*, 14: 253–267.

"Machines learning values" (2020), *The Ethics of Artificial Intelligence*, ed. S. Matthew Liao. New York: Oxford University Press.

"Composition as pattern" (2019), *Philosophical Studies*, 176(5): 1119–1139.

"Superintelligence as superethical" (2017), *Robot Ethics 2.0*, eds. Patrick Lin, Ryan Jenkins, and Keith Abney. New York: Oxford University Press.

"Is it good for them too? Ethical concern for the sexbots" (2017), *Sex Robots: Philosophical, Ethical, and Social Implications*, eds. John Danaher and Neil McArthur. Cambridge: MIT Press.

"A normative yet coherent naturalism" (2014), *Philo*, 17(1):77–91.

"Toward an algorithmic metaphysics" (2013), *Algorithmic Probability and Friends: Bayesian Prediction and Artificial Intelligence*, ed. David Dowe. Lecture Notes in Artificial Intelligence 7070:306–317.

"Utilitarian epistemology" (2013), *Synthese*, 190(6):1173–1184.

"Designing people to serve" (2011), *Robot Ethics: The Ethical and Social Implications of Robotics*, eds. Patrick Lin, Keith Abney, and George Bekey. Cambridge, MA: MIT Press.

"Analysis, schmanalysis" (2008), *Canadian Journal of Philosophy*, 38(2):289–300.

"The ethics of robot servitude" (2007), *Journal of Experimental and Theoretical Artificial Intelligence*, 19(1):43–54.

"Construing faith as action won't save Pascal's wager" (2006), *Philo*, 9(2):221–229.

"Functions, creatures, learning, emotion" (2004), *Architectures for Modeling Emotion: Cross-Disciplinary Foundations*, eds. Hudlicka, E. and Cañamero, L. The American Association for Artificial Intelligence, AAAI Press.

## EXTERNAL GRANTS

**Open Philanthropy** Grant for teaching reduction to support AI alignment research, 2025–2027.

**Survival and Flourishing** Grant for teaching reduction to support AI alignment research, fall 2024.

**Nonlinear / Center for AI Safety** General grant to support more AI alignment research on the margin, 2023–2024.

**The Long-Term Future Fund** One-course teaching reduction for each of two semesters 2022–2023, again to support more AI alignment research (in particular foundational approaches to the nature of *agency*).

**Survival and Flourishing** A supplementary semester of sabbatical, 2021–2022, to support further research on AI alignment (in particular formal models of value change).

## SELECTED PRESENTATIONS

"AI 2037: AI akrasia" (with Abram Demski, Sahil, and Tushita Jha) and "Optimization processes and agents who give a damn" (with Rosa Cao), ILIAD II: ODYSSEY Theoretical AI Alignment Conference at Lighthaven, Berkeley CA, fall 2025.

"Desire homuncularism: Agency, ethical standing, and skin in the game", AI & Humanity Lab, University of Hong Kong, spring 2025.

"AI sentience and AI agency", University at Buffalo Cognitive Science Colloquium, spring 2025.

"Abstraction by patterns", "The specter of teleology", "Learning your one true utility", ILIAD Theoretical AI Alignment Conference at Lighthaven, Berkeley CA, fall 2024.

"Abstraction for AI alignment", St. Joseph's University Conference in Honor of Todd Moody, Philadelphia PA, spring 2024.

Comments on Herman Cappelen's "Value alignment: what it is and why it's a dangerous idea", Pacific Division American Philosophical Association, Portland OR, spring 2024.

"From AI alignment to information theory", Computer Science and Engineering PhD Seminar, University of Louisville, spring 2024.

"AI and ChatGPT. Should we be worried?", National Association of Scholars online panel with Barry Smith and Jobst Landgrebe, spring 2023.

"Learning your One True Utility function", Metaethics and AI Workshop, Eindhoven University of Technology, summer 2022.

"Abstractions by patterns", Real Patterns Workshop, the Santa Fe Institute (virtual), spring 2022.

Comments on Alex Meehan's "Do we need a revolution? Two notions of (probabilistic) independence", Eastern Division American Philosophical Association (virtual), winter 2021.

"Should robots have rights?" and "Has technology replaced religion?" panels, Web Summit 2019 tech conference, Lisbon, Portugal, fall 2019.

"Machines learning values" (multiple presentations)

- Ethics and Artificial Intelligence, University of Agder (Norway), fall 2017.
- Carleton College invited talk, spring 2019.
- City of Good Neighbors Conference, Buffalo NY, spring 2019.
- The Center for Inquiry, Buffalo NY, summer 2019.
- Canisius College invited talk, fall 2019.
- SUNY Fredonia invited talk, spring 2021.

"Explanations as patterns" (multiple presentations)

- Real Patterns (Logos) invited talk, University of Barcelona, fall 2018.
- Scientific Explanations Competing and Conjunctive, University of Utah, summer 2019.

"Explanatory unification as information compression", 11th Munich-Sydney-Tilburg/Turin (MuST) conference (poster presentation), University of Turin, summer 2018.

"Superintelligence as superethical", Ethics of Artificial Intelligence, New York University, fall 2016.

"The statistical and unification approaches to explanation unified—statistically", Buffalo Logic Colloquium, spring 2016.

"Composition as pattern" (multiple presentations)

- Creighton Club, Syracuse University, fall 2015.
- Society for the Metaphysics of Science, Rutgers University, fall 2015.
- Society for Exact Philosophy, California Institute of Technology, spring 2014 (poster presentation).
- Buffalo Logic Colloquium, University of Buffalo, fall 2013.

"Explanation as information compression: Unificationism made precise", Algorithms and Complexity in Mathematics, Epistemology and Science, University of Western Ontario, spring 2015.

"Composition from life to organization: A slippery slope for van Inwagen", Society of Christian Philosophers, Eastern conference meeting, Niagara University, fall 2014.

"Is Rosenberg's scientism naturalistic? (An enchanted naturalism)", Ernst Mach Workshop on Disenchanted Naturalism, Department of Analytic Philosophy of the Institute of Philosophy, Academy of Sciences of the Czech Republic, spring 2014.

"Toward an algorithmic metaphysics", Solomonoff 85th Memorial Conference, Monash University, Australia (virtual), fall 2011.

"The ethics of human impairment" (with James Delaney), Human Enhancement Symposium, Center for Values in Medicine, Science, and Technology, the University of Texas at Dallas, spring 2011.

"Naturalism as a coherent ideology", University of Wisconsin-Milwaukee invited talk, fall 2009.

Comments on Carl Wagner's "Jeffrey conditioning and external Bayesianity", the Formal Epistemology Workshop, University of Wisconsin-Madison, summer 2008.

"Utilitarian epistemology and the value of knowledge", the Value of Knowledge Conference, Vrije Universiteit Amsterdam, summer 2007.

"Minimum message length as a truth-conducive simplicity measure", the Formal Epistemology Workshop, Carnegie Mellon University, spring 2007.

"Embodied intelligence without the bodies", the University of Buffalo Center for Cognitive Science, spring 2007.

"Naturalism is (literally) self-explanatory", the Center for Inquiry, spring 2007.

"The ethics of robot servitude" (multiple presentations)

- North American Computing and Philosophy Conference, summer 2006.
- University of Buffalo Center for Cognitive Science, fall 2006.
- Canisius College, spring 2007.

"Learning and computational epistemology", Western Michigan University invited speaker, fall 2005.

"Some future: comments on Gilbert Harman's *The future of the a priori*". University of Michigan, spring 2001 colloquium.

## SELECTED REFERENCES

**John Basl**

Northeastern University
j.basl|at|northeastern.edu

**David Chalmers**

New York University
chalmers|at|nyu.edu

**Abram Demski**

Machine Intelligence Research Institute
abram|at|intelligence.org

**John Wentworth**

Independent AI alignment researcher
jwentworth|at|g.hmc.edu