

DANIEL ALEXANDER HERRMANN

curriculum vitae

CONTACT INFORMATION

address Philosophy Department
 UNC Chapel Hill
 Caldwell Hall, CB# 3125
 240 East Cameron Ave.
 Chapel Hill, NC 27599-3125

email danher@unc.edu

website www.danielherrmann.ca

last updated: October 28, 2026

RESEARCH INTERESTS

AOS Philosophy of AI, Philosophy of Science, Game & Decision Theory

AOC Formal & Social Epistemology, Logic, Philosophy of Quantum Mechanics

ACADEMIC APPOINTMENTS

2025-present UNIVERSITY OF NORTH CAROLINA AT CHAPEL HILL
 Assistant Professor
 Department of Philosophy

2023-2025 UNIVERSITY OF GRONINGEN
 Postdoctoral Research Fellow
 Faculty of Philosophy

EDUCATION

PhD *2017-2023* UNIVERSITY OF CALIFORNIA, IRVINE
 Logic and Philosophy of Science
 Dissertation: *Naturalizing Decision Theory*
 Committee: Simon Huttegger (chair), Brian Skyrms,
 Jeffrey Barrett, Toby Meadows, James Joyce (external)

B.A.Sc. *2013-2017* QUEST UNIVERSITY CANADA
 Arts & Sciences
 Focus on Philosophy & Computer Science

BOOKS

- B1 **Herrmann, D.**, & Levinstein, B. *Radical AI Interpretability*, forthcoming with *Cambridge University Press*

DISCIPLINARY JOURNAL ARTICLES

- A10 Mohseni, A., **Herrmann, D.**, & Orona, G. (2026) “Why are the Social Sciences Hard? Two New Hypotheses” *Studies in History and Philosophy of Science*
- A9 **Herrmann, D.**, Chari, A., Qian, I., Sharvesh, S., & Levinstein, B. (2026) “A Decision-Theoretic Approach for Managing Misalignment” *Proceedings of the Second Conference of the International Association for Safe and Ethical Artificial Intelligence*
- A8 **Herrmann, D.**, & Rothfus, G. (2025) “Deference & Decision” *Theory and Decision*
- A7 **Herrmann, D.**, & Levinstein, B. (2025) “Standards for Belief Representations in LLMs” *Minds and Machines*
- A6 Levinstein, B., & **Herrmann, D.** (2025) “Still No Lie Detector for Language Models: Probing Empirical and Conceptual Roadblocks” *Philosophical Studies*
- A5 VanDrunen, J., & **Herrmann, D.** (2026) “Naturalizing Natural Salience”, *British Journal for the Philosophy of Science*
- A4 **Herrmann, D.** (2022) “Prediction with Expert Advice Applied to the Problem of Prediction with Expert Advice”, *Synthese*
- A3 **Herrmann, D.**, & VanDrunen, J. (2025) “Sifting the Signal from the Noise”, *British Journal for the Philosophy of Science*
- A2 **Herrmann, D.**, & Skyrms, B. (2025) “Invention and Evolution of Correlated Conventions”, *British Journal for the Philosophy of Science*
- A1 **Herrmann, D.** (2020) “PAC Learning and Occam’s Razor: Probably Approximately Incorrect”, *Philosophy of Science*

TEACHING & LEARNING JOURNAL ARTICLES

- T1 Orona, G., Pritchard, D., Arum, R., Eccles, J., Dang, Q., Copp, D., **Herrmann, D.**, Rushing, B., & Zitzmann, S. (2023) “Epistemic virtue in higher education: testing the mechanisms of intellectual character development”, *Current Psychology*

JOURNAL BOOK REVIEWS

- K2 Rushing, B., & **Herrmann, D.** (2025) “Review of Cameron J. Buckner’s From Deep Learning to Rational Machines: What the History of Philosophy Can Teach Us about the Future of Artificial Intelligence,” *Philosophy of Science*
- K1 **Herrmann, D.**, & Freeborn, D.P.W. (2022) “Wayne C. Myrvold. Beyond Chance and Credence: A Theory of Hybrid Probabilities,” *Philosophia Mathematica*

ARTICLES UNDER REVIEW

- R5 Saal, J., Shai, A., Levinstein, B., **Herrmann, D.**, & Riechers, P. “Transformers Develop Separable Posterior and Value Representations That Drive Choice” (Under review; equal senior authorship with P. Riechers)
- R4 **Herrmann, D.**, Mohseni, A., Levinstein, B., & Rushing, B. “A Bayesian Reduction of Causation” (Manuscript available)
- R3 Grote, T., **Herrmann, D.**, & Vernade, C. “Grounding Human-Machine Decision-Making” (Manuscript available)
- R2 Sterkenburg, T.F., **Herrmann, D.**, & Romeijn, J.-W. “Benign Interpolation and Occam’s Razor” (Manuscript available)
- R1 **Herrmann, D.** “Subjective Naturalism and Act Probabilities” (Under review)

WORKS IN PROGRESS

- W7 Cibinel, P., & **Herrmann, D.** “Intrapersonal Comparisons under Preference Uncertainty”
- W6 Rothfus, G., & **Herrmann, D.** “Completing Counterfactual Content”
- W5 Rothfus, G., & **Herrmann, D.** “Causal Decision Theory Without Possible Worlds”
- W4 **Herrmann, D.** “The End of Deliberation”
- W3 **Herrmann, D.** “Decision Under Certainty”
- W2 Garrabrant, S., **Herrmann, D.**, & Lopez-Wild, J. “Cartesian Frames” (Manuscript available)
- W1 **Herrmann, D.** “Measure for Measure: Problems with Probability in Decohering Worlds” (Manuscript available)

PRESENTATIONS († ≡ invited, * ≡ poster)

- P37 *Panel: The Language of AI Safety for Science*
†AISciK Workshop, NeurIPS · Atlanta 2026
- P36 *A Bayesian Reduction of Causation*
Philosophy of Science Association Biennial · San Diego 2026
Philosophy of Science Meets Machine Learning · Munich 2026
- P35 *Radical AI Interpretability*
†Minds, Meaning, & Machines · Hong Kong 2026
AI Agency & Interpretability Workshop · Pittsburgh 2026
- P34 *Margins of Misalignment: Delegation to AI Beyond the Bayesian Idealization*
PPE Society Annual Meeting · New Orleans 2026
- P33 *Why are the Social Sciences Hard?*
PPE Society Annual Meeting · New Orleans 2026
- P32 *Intrapersonal Comparisons under Preference Uncertainty*
PPE Society Annual Meeting · London 2026
Meeting of the Society for Social Choice and Welfare · Tokyo 2026
- P31 *Can We Attribute Beliefs and Desires to AIs? A Radical Approach*
†Bennington Center for AI Speaker Series · Bennington 2026
- P30 *The Metaphor Problem for Belief Representation in Large Language Models*
†Pacific APA Meeting · Online 2026
- P29 *A Decision-Theoretic Approach for Managing Misalignment*
Intl. Assoc. for Safe and Ethical AI · Paris 2026
- P28 *Machine Minds and the Propositional Projection Problem*
†Central APA Meeting · Chicago 2026
- P27 *Comments on Cappelen and Dever's "Making AI Intelligible"*
†Eastern APA Meeting · Baltimore 2026
- P26 *Splitting the Atom: Atomless Semantics for Counterfactuals*
Philosophy of Science Association Biennial · San Diego 2026
PPE Society Annual Meeting · New Orleans 2025
- P25 *Naturalizing Causation in Causal Models*
†Agent Foundations Workshop · Pittsburgh 2025
- P24 *Pragmatics, Empiricism, and Self-Assembling Games*
†Self-Assembling Games Workshop · Irvine 2025
- P23 *Deference and Decision*
†Agent Foundations Workshop · Pittsburgh 2025
†Pittsburgh Formal Epistemology Workshop · Pittsburgh 2025
†UCI LPS Game Theory Lab · Irvine 2025

PRESENTATIONS († ≡ invited, * ≡ poster)

- P22 *Manageable Misalignment*
†Agent Foundations Workshop · Pittsburgh 2025
PPE Society Annual Meeting · New Orleans 2024
- P21 *Standards for Belief Representation in LLMs*
†Bennington Science Workshop · Bennington 2026
†MCMP Research Seminar in Decision and Action Theory · Munich 2024
Philosophy of Science Association Biennial · New Orleans 2024
†InDeep Consortium Meeting · Groningen 2024
Philosophy of Science Meets Machine Learning · Tübingen 2024
- P20 *Decision Theory Foundations*
†ILIAD Conference · Berkeley 2024
- P19 *Measuring the Beliefs of LLMs*
†Purdue University Philosophy Colloquium · West Lafayette 2024
†AI & Logic Talk · Groningen 2024
- P18 *Interpersonal Comparisons of Utility: A Conventional Take*
Hannover-Groningen Workshop · Hannover 2024
Philosophy of Science Meets Machine Learning · Tübingen 2023
- P17 *Probing the Beliefs of Large Language Models*
†Philosophy, Sciences, and Humanities Lecture · Groningen 2023
- P16 *From Savage to Jeffrey-Bolker*
†Agent Foundations Workshop · Oxford 2023
- P15 *Endogenizing Control*
Issues in Dynamic Decision Theory Workshop · Konstanz 2023
- P14 *Artificial Agency*
†Epistemology and Theory of Machine Learning · Munich 2023
- P13 *Attention!*
†University of Groningen Faculty of Philosophy · Groningen 2023
†Santa Fe Institute Research Seminar · Santa Fe 2023
- P12 *Deciding as Learning*
†Seaman Lecture, University of Idaho · Moscow 2023
†Cambridge Philosophy Faculty · Cambridge 2022
- P11 *Naturalizing Natural Salience*
Philosophy of Science Association Biennial · Pittsburgh 2022
- P10 *Measure for Measure*
†California Quantum Interpretation Network · Chapman University 2022
- P9 *Jeffrey-Bolker Decision Theory and Embedded Agency*
†Human Aligned AI Summer School · Prague 2022
- P8 *Ramification Functions*
*Formal Epistemology Workshop · Irvine 2022

PRESENTATIONS († ≡ invited, * ≡ poster)

- P7 *Sifting the Signal from the Noise*
ICLR Emergent Communication Workshop · Online 2022
- P6 *Rationality for Real Agents*
Philosophy of Science Association Biennial · Baltimore 2021
- P5 *Why Randomize, Really?*
Randomization Workshop · Online 2021
- P4 *The Problem of Self-Knowledge: Walking the Tightrope*
†Logic Research Seminar, Charles University · Online 2020
- P3 *PAC Learning and Occam's Razor*
European Philosophy of Science Association Biennial · Geneva 2019
†University of Groningen Faculty of Philosophy · Groningen 2019
*Philosophy of Science Association Biennial · Seattle 2018
- P2 *Bars and Belief: The Probability Puzzle Persists*
IMBS Luce Graduate Student Conference · Irvine 2019
Our Futures Conference · Squamish 2019
- P1 *Inventing Correlated Conventions*
Pacific APA · Online 2021
IMBS Seminar · Irvine 2019

INTERNSHIPS

BenchSci Machine Learning Summer Intern · 2016

AWARDS & FELLOWSHIPS

Principles of Intelligence Horizon Scanning Research Grant · 2026

Long-Term Future Fund Research Grant · 2022-2023

Principles of Intelligent Behavior in Biological and Social Systems Research Fellowship · 2022

Outstanding Teaching Award · 2022

Cascade Mentorship Fellow · 2021 - 2023

History and Philosophy of Science Travel Award · 2022

Associate Dean Fellowship · 2021

Machine Intelligence Research Institute Summer Fellowship · 2018

UCI Summer Research Fellowships · 2017-2021

UCI Merit Scholarships · 2017-2023

TEACHING EXPERIENCE

AI Interpretability	Instructor	UNC	Spring 2026
AI & The Future of Humanity	Instructor	UNC	Spring 2026
Dissertation Research Seminar	Instructor	UNC	Fall 2025
AI & The Future of Humanity	Instructor	UNC	Fall 2025
Rationality & Decision	Instructor	UG	Spring 2025
Philosophy of Machine Learning	Instructor	UG	Fall 2024
Ethics & Epistemology of AI	Instructor	UG	Fall 2023
Basic Economics	Teaching Assistant	UCI	Fall 2022
Critical Reasoning	Instructor	UCI	Winter 2022
Probability and Statistics	Teaching Assistant	UCI	Fall 2021
The Good Life	Teaching Assistant	UCI	Spring 2020
Probability and Statistics	Teaching Assistant	UCI	Winter 2020
Naturalized Epistemology	Teaching Assistant	UCI	Fall 2019
Probability and Statistics	Teaching Assistant	UCI	Spring 2019
Naturalized Epistemology	Teaching Assistant	UCI	Fall 2018
Intro to Inductive Logic	Teaching Assistant	UCI	Spring 2018
Critical Reasoning	Teaching Assistant	UCI	Winter 2018

PROFESSIONAL SERVICE

Referee (Journals): *Australasian Journal of Philosophy, Economics and Philosophy, Episteme, Erkenntnis, European Journal for the Philosophy of Science, Foundations of Physics, Journal for General Philosophy of Science, Mind, Philosophy and the Mind Sciences, Philosophy Compass, Philosophy of Science, Philosophical Studies, Rationality and Society, Synthese, The British Journal for the Philosophy of Science, The Review of Symbolic Logic, Theory and Decision, Ethics and Information Technology, Harvard Data Science Review*

Referee (Conferences): International Association for Safe and Ethical AI (2027), Formal Epistemology Workshop (2022)

Organizer: Philosophy of AI Workshop, UNC Chapel Hill (2027); Workshop on AI Agency, Carnegie Mellon University (2026, with A. Mohseni)

Department: Faculty Search Committee, Philosophy of AI, UNC (2026-27); Core Faculty, Philosophy, Politics & Economics Program, UNC (2025-present)

Coordinator: ILIAD Agent Foundations Track Lead (2024), Formal Epistemology Reading Group (2019-2022), Graduate Representative (2020-2021), Work in Progress Graduate Student Seminar (2019)
Support Staff: Formal Epistemology Workshop (2022)
Training Facilitator & Panelist: Teaching Assistant Professional Development Program (2019)
Mentor: Student Mentoring, UNC (2025-present); Iliad Fellowship Mentor (2026); PIBBSS Fellowship Mentor (2026, co-mentor with P. Riechers); Algoverse AI Safety Fellowship Mentor (2025, with B. Levinstein); UG Master's Thesis Supervision (2024), UG Bachelor's Thesis Supervisor (2023-2024), UCI Cascade Mentor for Graduate Students (2021-2023)

SELECTED GRADUATE COURSEWORK († ≡ directed study, * ≡ audit)

<i>Rational Choice, Decision/Game Theory, & Formal Epistemology</i>	Bayesianism	B. Skyrms & S. Huttegger	UCI	2022
	Deliberational Decision Theory	S. Huttegger	UCI	2021
	*Decision Theory	J. Joyce	U-M	2020
	Signaling Games	B. Skyrms	UCI	2019
	Utilitarianism & Measurement Theory	B. Skyrms & L. Narens	UCI	2019
	Bayesian Epistemology	S. Huttegger	UCI	2018
	Chance	B. Skyrms	UCI	2017
<i>Philosophy of Science & Mathematics</i>	Logical Empiricism	B. Skyrms & S. Huttegger	UCI	2023
	Classical Field Theory II	J. Weatherall	UCI	2021
	Classical Field Theory I	J. Weatherall	UCI	2021
	†Leibniz's Philosophy of Probability	S. Huttegger	UCI	2020
	Foundations of Quantum Mechanics	J. Barrett	UCI	2018
	Social Dynamics	B. Skyrms & L. Narens	UCI	2018
	Evolution of Morality	K. Stanford	UCI	2018
	Wittgenstein's Philosophy of Logic	P. Maddy	UCI	2018
	Probability & Determinism	J. Manchak	UCI	2017
	Philosophy of Math	P. Maddy	UCI	2017
<i>Logic/Mathematics</i>	Descriptive Set Theory	T. Meadows	UCI	2023
	Category Theory	T. Meadows	UCI	2021
	Set Theoretic Geology	T. Meadows	UCI	2020
	Set Theory II	M. Zeman	UCI	2020
	Set Theory I	M. Zeman	UCI	2019
	Large Cardinals and Determinacy	T. Meadows	UCI	2019
	Analysis II	P. Ivanisvili	UCI	2018
	Analysis I	P. Ivanisvili	UCI	2018
	Incompleteness and Undecidability	J. Schatz	UCI	2018
	Metalogic	E. Galebach	UCI	2018
	Set Theory	S. Walsh	UCI	2017