PHILIPP KORALUS

McCord Professor of Philosophy and AI
University of Oxford

philipp.koralus@stcatz.ox.ac.uk www.koralus.net

APPOINTMENTS

MCCORD PROFESSOR OF PHILOSOPHY AND AI
Institute for Ethics in AI — from Aug 2024
University of Oxford

FULFORD CLARENDON PROFESSOR OF PHILOSOPHY AND COGNITIVE SCIENCE
Faculty of Philosophy
St. Catherine's College — 2023-24
University of Oxford

ASSOCIATE PROFESSOR
Faculty of Philosophy
St. Catherine's College — 2013-23
University of Oxford

JAMES S. MCDONNELL POSTDOCTORAL FELLOW,
Philosophy-Neuroscience-Psychology Program
Washington University in St. Louis — 2011-2013

FELLOW IN THE NOTRE DAME INSTITUTE FOR ADVANCED STUDY,
University of Notre Dame — 2010-2011

VISITING FELLOW
Munich Center for Mathematical Philosophy,
Ludwig-Maximilians-Universität, Munich — 2012; 2013

VISITING RESEARCHER,
Institut Jean-Nicod,
École Normale Supérieure, Paris — 2011

EDUCATION

Ph.D. Philosophy and Neuroscience, Princeton University — 2010
B.A. Philosophy, Pomona College — 2005

BOOKS

Koralus, P. (2023). Reason and Inquiry: The Erotetic Theory. Oxford University Press.

Abstract: Reason and Inquiry presents a unified theory of the human capacity for reasoning and decision-making. The erotetic theory accounts for a diverse range of empirically documented fallacies and framing effects and shows how the same mental processes that yield fallacies can yield what logicians call first-order validity and probabilistic coherence in reasoning, as well as rational decision-making as conceived by economists. The central idea is that our minds naturally aim at resolving issues, and if we are sufficiently inquisitive in the process, we can avoid mistakes. The erotetic theory holds that both the successes and the failures of reason are due to this aim. Rationality is secured if we reach erotetic equilibrium, which is proved as a theorem.

Reviews of Reason and Inquiry:

"It is easy for researchers in Artificial Intelligence (AI) to get excited with our technical achievements and lose track of the big questions: what is intelligence, and how does it work? This thought provoking and wide-ranging book prompts us to look again at our field: to revisit the most basic questions surrounding our endeavour, and, perhaps most importantly, to consider new directions for the future."

-Michael Wooldridge, Co-Director for Artificial Intelligence at The Alan Turing Institute

"Philipp Koralus presents a bold, unified, and original theory of human reason, as centred on asking and answering questions. He backs it up with a powerful combination of experimental evidence and logical analysis. Reason and Inquiry is a major contribution to the philosophy of mind, the psychology of reasoning, and cognitive science, with implications for linguistics, epistemology, and decision theory. The erotetic theory looks set to be a key player in future debates on the nature of rationality."

-Timothy Williamson, FRSE, FBA, Wykeham Professor of Logic, University of Oxford

"An insightful treatment of reason and rationality, explaining many puzzles and integrating many viewpoints."

-Steven Pinker, Johnstone Professor of Psychology, Harvard University

PAPERS

1. Womersley, K., Koralus, P., Fulford, B., Handa, A., Peile, E. (2023). "Hearing the Patient's Voice in AI-enhanced Healthcare." *British Medical Journal (BMJ)*.
2. Koralus, P., Wang-Mascianica, V. (*under review*). "Humans in humans out: On GPT converging toward common sense in both success and failure."
3. Madsen, J., Carrella, E., Bailey, R., Koralus, P. (*2020*). "From reactive toward anticipatory fishing agents." *Journal of Simulation*. https://doi.org/10.1080/17477778.2020.1742588
4. Burgess, M. Drexler, M., Axtell, R., *et al.,* Koralus, P., *et al*. (2020). "Opportunities for agent-based modeling in human dimensions of fisheries." *Fish and Fisheries*. 21(3), p.570-587.
5. Madsen, J., Bailey, R., Carrella, E., Koralus, P. (2019*).* "Analytic versus computational cognitive models: Agent-based modeling as a tool in cognitive sciences." *Current Directions in Psychological Science,* 28(3), pp. 299-305.
6. Koralus, P. and Mascarenhas, S. (*2018*). "Illusory inferences in a question-based theory of reasoning." In: Horn, L. and Turner, K. (Eds.) *An Atlas of Meaning (Current Research in the Semantics/Pragmatics Interface*). Brill.
7. Koralus, P. and Alfano, M. (2017). "Reasons-based moral judgment and the erotetic theory." In: Bonnefon. J.-F. Tremoliere, B. (Eds.). *Moral Inferences*. Psychology Press: Routledge.
8. Mascarenhas, S., Koralus, P. (2016). "Illusory inferences with quantifiers." *Thinking & Reasoning*, 23(1), pp. 33-48.
9. Koralus, P. (2016) "Can visual cognitive neuroscience learn anything from the philosophy of language? Ambiguity and the topology of neural network models of multistable perception." *Synthese, 193(5), pp. 1409-32.*
10. Parrott, M. and Koralus, P. (2015) "The erotetic theory of delusional thinking." *Journal of Cognitive Neuropsychiatry.* 20(5), pp. 398-415. (commentary by John McKay and symposium on http://imperfectcognitions.blogspot.co.uk)
11. Mascarenhas, S. and Koralus, P. (2015). "Illusory inferences: disjunctions, indefinites, and the erotetic theory of reasoning." In: Noelle, D. C., et al. (Eds.). *Proc. 37th Meeting of the Cognitive Science Soc*, Cognitive Science Society.
12. Koralus, P. "The erotetic theory of attention: questions, focus, and distraction." (2014). *Mind & Language*, 29(1), pp. 26-50. (selected for symposium at www.philosophyofbrains.com, replies by Chris Mole, Felipe de Brigard, Sebastian Watzl, et al.)
13. Koralus, P. (2014). "Attention, Consciousness, and the semantics of questions." *Synthese*, 191(2), pp. 187-211.
14. Koralus, P. and Mascarenhas, S. (2013). "The erotetic theory of reasoning: Bridges between formal semantics and the psychology of propositional deductive inference." *Philosophical Perspectives*, 27, pp. 312-365.
15. Koralus, P. (2013). "Descriptions, ambiguity, and representationalist theories of interpretation." *Philosophical Studies*, 162(2), pp. 275-290.
16. Koralus, P. (2013). Review of Jonathan Berg's *Direct Belief. Notre Dame Philosophical Reviews.*
17. Koralus, P. (2012). "The open instruction theory of attitude reports and the pragmatics of answers." *Philosopher's Imprint*, Vol. 12(4).

18. Parkinson, C. Sinnott-Armstrong, W., Koralus, P., Mendelovici, A., McGeer, V. and Wheatley, T. (2011). "Is morality unified? Evidence that distinct neural systems underlie moral judgments of harm, dishonesty, and disgust." *Journal of Cognitive Neuroscience,* Vol. 23(10), pp. 3162-80.

RESEARCH GRANTS

1. HAI Lab starting grant. PI: Philipp Koralus. £500,000. Cosmos Institute. Start date August 2024.
2. AI and human flourishing. PI: Philipp Koralus. £6,283. Cosmos Institute. Start date November 2023.
3. Computational foundations for the erotetic theory of reason. PI: Philipp Koralus, Sean Moss (Computer Science). £43,200. John Fell Fund. Start date July 2023.
4. Book manuscript support grant. PI: Philipp Koralus. £10,412. Oxford Humanities Division. Start date October 2017.
5. Reasoning with probabilities. PI: Philipp Koralus. £10,000. Laces Trust. Start date November 2017.
6. Mitigating delusional conclusions. PI: Philipp Koralus. £24,047. John Fell fund. Start date July 2016.
7. The Erotetic Theory of Reasoning. PI: Philipp Koralus. £153,482. Laces Trust. Start date: September 1st 2014-May 2017.
8. Project Director for Cognition and Decision-Making within Oxford Martin School / Ocean Conservancy project on marine ecology. PI: Richard Bailey (Oxford Geography). £1,500,000. Start date: January 2016.
9. A question-based approach to reasoning and decision-making. PI: Philipp Koralus. £6,251. John Fell Fund. Start date: January 2015.
10. Starting Grant for the Laboratory for the Philosophy and Psychology of Rationality and Decision. PI: Philipp Koralus. £4,000. Laces Trust. Start date: September 1st 2013.

Total grants as PI: £757,675

POST-DOC AND GRADUATE RESEARCH SUPERVISION

Beau Mount (Post-doctoral Fellow)
Now Associate Professor of Logic at New College, University of Oxford 2016-17

Jens Madsen (Post-doctoral Fellow, co-advised with Richard Bailey)
Now Assistant Professor of Psychology LSE 2016-18

Salvador Mascarenhas (Post-doctoral Fellow)
Now Full Professor of Cognitive Science, École Normale Supérieure, Ulm, Paris 2014-16

Benjamin Lang (DPhil advisee – Philosophy – *ongoing*)
Dissertation: "Responsibility Gaps in the Ethics of AI" 2023-

William Wells (DPhil advisee – Philosophy – *ongoing*)
Dissertation: "The Cognitive Science of Moral Judgment" 2020-

Vincent Wang (DPhil co-advisee – Department of Computer Science, now Researcher at Quantinuum)
Dissertation: "String Diagrams for Text" 2020-23

Raphaël Millière (DPhil co-advisee – Philosophy – now Assistant Professor Macquarie University)

Dissertation: "A Pluralist Account of Self-Consciousness" 2017-19

Salvador Gouveia (BPhil Thesis supervision – Philosophy – *Thesis distinction awarded by Examiners*)
Thesis: "Creativity versus Conservatism in Science: A Simulation-based Investigation" 2022

Anton Fedotov (MCompSciPhil Thesis supervision)
Thesis: "On the Content of Consciousness in Humans and Machines." 2024

Xin Guan (MMathPhil Thesis supervision)
Thesis: "Is AI Intelligent because of its Instrumental Rationality?" 2023

Tianjin Li (MMathPhil Thesis supervision)
Thesis: "Predictive Processing: Commitments, Mental Functions, and Illusions" 2023

INVITED TALKS

1. "The erotetic theory of reason," Topos Institute, Berkeley, August 2023
2. "How to be a Mental Model radicalist," Conference in Honor of Philip Johnson-Laird, University College London, July 2023.
3. "A question-based theory of human reason and a benchmark for GPT," University of Bochum, June 2023
4. "A question-based theory of human reason and a benchmark for LLMs," Imbue (Generally Intelligent), San Francisco, May 2023
5. "A question-based theory of human reason and a benchmark for GPT (with V. Wang)," Language group, Stanford U, May 2023.
6. "A question-based theory of human reason and a benchmark for GPT," Kathy Wilkes Memorial Conference, April 2023.
7. "The erotetic theory of attention," Friends of Attention Meeting, Princeton University, May 2022.
8. "Reasoning and consciousness." Keynote lecture at iCog 5: Interdisciplinary Approaches to Higher Cognitive Function, University of Reading, February 2019.
9. "An erotetic account of scalar implicature." (with Vincent Wang). École Normale Supérieure, Paris, June 2018.
10. "Ordinary reasoning with arbitrary objects", Kit Fine's seminar on arbitrary objects, NYU, February 2018.
11. "How can buggy minds carry us to the moon?" *The Wotton's Society*, Eton College, UK Nov. 9, 2017.
12. "Reasoning as inquiry." Keynote Lecture at Poznan Reasoning Week, University of Poznan, Poland. Sept. 10th, 2016.
13. "Some reflections on mental models: mind, language, and society." Vittorio Girotto Memorial Conference, Birkbeck College, London, July 27th, 2016.
14. "The erotetic theory" Two lectures at University of Barcelona/LOGOS, May 4th -5th, 2016.
15. "The Erotetic Theory as a Unified Approach to Reasoning, Judgment, and Decision-Making." London Judgment and Decision Making seminar, UCL, June 8th, 2016.
16. "Questions, Reasons, and Fish." Ocean Conservancy POSEIDON meeting, New Orleans, April 13, 2016.
17. "New Fallacies of Reasoning." Philosophy of Language Group. UCL. March 10th, 2015.
18. "Questions and the Norms of Reason." Norms of Inquiry Workshop. NYU. Feb. 12th, 2015.
19. "Decisions, Questions, and Illusory Reasons." Choice Group. LSE. December 11th, 2013.
20. "Questions, Cognition, and Conditionals: Bridging the Gap Between Psychology and Formal Semantics." The New York Philosophy of Language Workshop, NYU, February 25th 2013.
21. "Attention, Meditation, and Consciousness." Eastern APA, Atlanta, December 2012.
22. "The Erotetic Theory of Reasoning." Munich Center for Mathematical Philosophy, June 6th, 2012.
23. "From the Erotetic Theory of Attention: Toward a General Theory of Prefrontal Cortex Function." Cognitive Control and Psychopathology Laboratory, Washington University in St. Louis, February 20th, 2012.
24. "Consciousness and the Erotetic Theory of Attention." University of Illinois, Urbana-Champaign, January 26th, 2012.
25. "Perception, Action, and the Semantics of Questions." University of Connecticut, January 23rd, 2012.
26. "The Erotetic Theory of Attention." Workshop on the philosophy of attention (with Jesse Prinz), Institut Jean-Nicod, CNRS/École Normale Supérieure, Paris, June 15th, 2011.
27. "The Erotetic Theory of Attention." Philosophy Department, Washington University in St. Louis, May 4th, 2011.
28. "Against the Received View of *de re*/*de dicto* Ambiguities." MERG Experimental Philosophy Lab, The City University of New York, November 5th, 2010.
29. "The Disunity of Neural Correlates of Moral Judgment." Morality and Cognition Workshop, Notre Dame University, October 8th, 2010.
30. "The Open Instruction Theory of Attitude Reports." Cognitive Science Symposium, Dept. of Philosophy, The Graduate Center, The City University of New York, October 23rd, 2009.

31. "On the Cognition and Acquisition of Sentences Reporting on Beliefs and Desires." Psycholinguistics and Cognition Lab, Princeton University, December 1st, 2008.
32. "New Views of the Necker Cube: Vision from a Linguistic Perspective." Robert Efron Lecture in Linguistics and Cognitive Science, Pomona College, April 10th, 2008.
33. "Toward Semantics and Pragmatics in Visual Feature Binding Theory: An Analysis of Necker Cubes and Duck Rabbits." Cognitive Science Symposium, Dept. of Philosophy, The Graduate Center, The City University of New York, December 7th, 2007.

CONFERENCE PRESENTATIONS

35. "Toward a content-based theory of reasoning and general intelligence." Int. Conference on Thinking, Paris, 2021 (with Sean Moss and Vincent Wang).
36. "A cognitive realistic model of decision-making in ocean ecology." Cognitive Science Society Meeting, Philadelphia, 2016.
37. "Free-form response vs. yes/no-question methodologies in human reasoning." Cognitive Science Soc., Philadelphia, 2016.
38. "Reasons-based moral judgment and the erotetic theory." Society for Philosophy and Psychology Meeting, University of Texas at Austin, TX, 2016.
39. "Illusory inferences from disjunctions and quantifiers: The erotetic theory" European Cognitive Science Soc. Meeting 2015, Turin.
40. "Illusory inferences and the erotetic theory." London Reasoning Workshop 2015, Birkbeck College, London August, 2015.
41. "Illusory inferences: disjunctions, indefinites, and the erotetic theory of reasoning" with Salvador Mascarenhas. Cognitive Science Society Meeting 2015, Pasadena, CA.
42. "Reasoning with quantifiers beyond syllogisms: illusory inferences with indefinites and the erotetic theory of reasoning." with Salvador Mascarenhas. 2015 Society for Philosophy and Psychology Meeting, Duke University.
43. "The Erotetic Theory of Reasoning." London Reasoning Workshop 2014, Birkbeck College, London July 9th, 2014.
44. "Decisions, Questions, and Illusory Reasons." Faculty of Philosophy, Oxford, March 7th, 2014.
45. "Attention, Questions, and Consciousness." Association for the Scientific Study of Consciousness Conference, Brighton, UK, June 5th, 2012.
46. "Questions Make us Rational." Washington University in St. Louis, April 26th, 2012.
47. "The Erotetic Theory of Attention: Focus, Questions, and Representationalism." Southern Society for Philosophy and Psychology Conference, March 22, 2012.
48. "The Open Instruction Theory of Attitude Reports." Public Dissertation Talk, Princeton University, May 20th, 2009.
49. "Comment on Rachel Sterken, 'Generics, Semantic Blindness and Mosquitoes'." Princeton-Rutgers Graduate Conference in Philosophy, Princeton University, March 28th, 2009.
50. "Comment on Shaun Nichols, 'Emotions, Norms, and the Moralization of Fairness'." Princeton Moral Psychology Conference, Princeton University, November 14th, 2008.
51. "Attitude Reports and Cognition." The ANU Philosophy Society, Research School of Social Sciences, Australian National University, August 5th, 2008.
52. "Attitude Reports Without Ambiguity and Hidden Indexicality." Annual Conference of the Australasian Association of Philosophy, La Trobe University, July 7th, 2008.

SELECTED TEACHING AT THE UNIVERSITY OF OXFORD

`Graduate Seminars

- Philosophy, AI, and Innovation (w/ Brendan McCord)
- Topics in Minds and Machines: Perception, Cognition, and ChatGPT (w/ Will Davies)
- A Theory of Reason: Philosophy, Psychology, and Algorithms. (w/ Sean Moss, Computer Science)
- The Philosophy and Psychology of Reasoning and Decision-making.
- The Philosophy of Cognitive Science. (w/ Martin Davies).
- Philosophical Foundations of Psychology. (for Msc Program in Psychology)
- The Philosophy Consciousness and Attention. (w/ Ian Phillips).
- Topics in Mind and Language. (w/ Salvador Mascarenhas, Linguistics).

Undergraduate

- Philosophy of Cognitive Science
- Philosophy of AI
- Philosophy of Mind

- Ethics
- Introduction to Moral Philosophy
- Introduction to General Philosophy

SERVICE

- Chair of the Board of Examiners, BPhil program in philosophy. MT 2017-2019
- Chair of the Board of Prelim Examiners, Philosophy, Politics, and Economics (PPE), 2021-22
- Faculty Board Member, Faculty of Philosophy, MT2019-TT23
- Member of Governing Body, St. Catherine's College, MT 2013-
- Course Coordinator, Msc Philosophical Foundations of Psychology, Psychology Department, Oxford, 2017-2018
- Director of Undergraduate Studies in Philosophy, Politics, and Economics (PPE), St. Catherine's College, 2016-2018, 2020-
- Lead Undergraduate Admissions Tutor, Philosophy, Politics, and Economics (PPE), St. Catherine's College, 2014, 17, 20, 22
- Finance Committee, St. Catherine's College, 2018-21
- Graduate Placement Officer, Faculty of Philosophy, 2017-2018
- Selection Committee / Board of Electors,
    - Early Career Fellowship in AI and Ethics (Faculty of Philosophy), Oxford, 2023
    - Associate Professorship in AI and Ethics (Faculty of Philosophy), Oxford, 2022
    - Associate Professorship in AI and Ethics (Faculty of Philosophy and Dept. of Computer Science), Oxford, 2020
    - Associate Professorship in AI and Ethics (Faculty of Philosophy), Oxford, 2020
    - Professorship in Philosophy of Language, Oxford, 2014-15
    - Statutory Professorship in Translational Cognitive Neuroscience, Oxford, 2014
    - Associate Professorship in Economics, Oxford, 2016-17
    - Visiting Fellows Selection Committee, St. Catherine's College, MT 2013-2018

- Refereeing

    Mind, Philosophical Studies, Synthese, Australasian Journal of Philosophy, Cambridge University Press, American Phil. Quarterly, Behavioral and Brain Sciences, Journal of Experimental Psychology, Cognitive Psychology, Memory & Cognition, Brain and Cognition, Journal of Philosophical Psychology, European Summer School of Logic, Language and Information, Springer Studies in Brain and Mind, University of Notre Dame Press, Minds and Machines, *Dialektica, The Review of Philosophy and Psychology, Acta Psychologica. European Research Council. Austrian Science Foundation. Deutsche Forschungsgemeinschaft*