Joel Artturi Saarinen

Toulouse, France | Joel-Artturi.Saarinen@irit.fr | Website | Github

Personal Details

- Born: 15 July 1998, Helsinki
- Languages: Finnish (Native/C2), English (Native/C2), French (C1), Italian (A2)

Research Statement

My research sits at the intersection of philosophical logic and machine learning. More specifically, I'm interested in how normative knowledge (what agents ought to do, value, or believe) can be formally represented and learned, and what distinguishes this from the learning of empirical facts. This includes work on deontic logic, learning theory, and multi-agent systems, with an eye toward what these frameworks imply for how AI systems and collective human institutions come to act on values.

Research Interests

- **Learning Theory** | Algorithmic LT · Machine LT · Formal/Logic-based LT · Bayesianism · Foundations of Probability
- **Information Theory** | Algorithmic IT · Algorithmic Randomness
- **Philosophical Logic** | Formal theories of truth · Deontic Logic · Formal Ethics
- **Game Theory** | Epistemic Game Theory, Logics for Multi-Agent Systems, Model-Checking in Games
- **Programming Languages** | Python, Haskell, ProbLog

Education

PhD **IRIT / L'Université de Toulouse III**, Logic — Oct 2025 – Sept 2028
- Thesis written as part of CaRE project in LILaC team under the supervision of Emiliano Lorini.

MSc. **University of Amsterdam**, Logic — Sept 2022 – June 2025
- Philosophy track with overall focus on Philosophy and Computation themes.
- **Featured Coursework:** Dynamic Epistemic Logic, Topology Logic and Learning, Bayesian Epistemology, Philosophical Logic, Modal Logic, Category Theory, Recursion Theory

BA **Occidental College**, Joint Bachelors Degrees in Philosophy and Mathematics, Minor in Cognitive Science — September 2017 – May 2021
- First Class Honours (UK scale)

Relevant Work Experience

Machine Intelligence Research Institute, Research Intern — Berkeley, CA, US — June 2022 – September 2022
- Worked as a generalist researcher during the summer trying to discover links between human intelligence and agency, making connections to human value/normative information learning problem. This involved a major literature review out, an eventual analysis, and a final output (found below under "Projects" section).

University of Southern California, Research Intern — Los Angeles, CA, US — July 2018 – August 2018
- Worked with Elsi Kaiser (Linguistics/Cognitive Science Dept.) and organized web-scraped linguistic data from Finnish expats living in Southern California with Google Sheets and Excel
- Developed extensive questionnaires based on data using Qualtrics meant to gauge language retention ability of subjects and summarized findings in written report

Master's Thesis

Limits of Solomonoff Induction Link:, June 2025

Supervisors: Francesca Zaffora Blando and Aybüke Özgün

My masters thesis examined the limits of Solomonoff induction as a formal theory of learning. Building on Hutter's result that the Solomonoff inductor fails to converge on data streams that are typical under the Lebesgue measure, I extended this non-convergence result to arbitrary computable measures satisfying a boundedness condition, strengthening the case that Solomonoff induction falls short as a theoretical ideal. The thesis closes with a philosophical discussion of what these impossibility results imply for the prospects of finding any ideal inductive method.

Projects

Bridging Statistical Learning Theory and Dynamic Epistemic Logic Link:, 2024

- Co-authored (with Paulius Skaisgiris) a work connecting fundamental formal notions of learning from statistical learning theory to dynamic epistemic logic using formal learning theory as a bridge. Assisted by Alexandru Baltag.

Pluralistic Bayesian Truth Convergence Link:, 2025

- Undertook an independent investigation of how Bayesianism could be adapted to allow updates for propositions with more normative content, breaking from tradition of considering seemingly only more empirical propositions. I propose some initial criteria for modifications and implement them. Supervised by Tom Schoonen.

Value Learning In The Absence Of Ground Truth Link:, 2023

- Secured funding for and worked independently on project examining various existing learning/decision-theoretic frameworks for learning normative truths and evaluating their effectiveness with respect to enabling AI systems to learn human values. In progress of implementing these frameworks in AI systems and running experiments on which one performs best according to desired criteria. Report reached front page of LessWrong forum. Supervised by Daniel Herrmann.

What Created Human Dominance? Link:, 2022

- Undertook summer research project figuring out what aspect of human cognition or collective intelligence allowed humans to become so dominant. Connections made problem of human value learning problem for AI systems made throughout. Supervised by Ronny Fernandez, Rick Korzekwa, and Katja Grace.

Talks

Theories of Learning (with Paulius Skaisgiris) May 2024

Cool Logic, University of Amsterdam/ILLC

Workshops/Conferences

Summer School on Ethical Design for AI, Lecce (Italy) September 2025
NiHiL Workshop, University of Amsterdam January 2024

Teaching

5354PHSC6Y: Philosophy of Science, Teaching Assistant Fall 2024

Graded assignments and exams and organized and ran tutorials for reinforcing class material for the Philosophy of Science course intended for masters students at the University of Amsterdam. Instructors for the course were Sebastian De Haro

Ollé and Eline de Jong.

Grants

AI Safety Research Grant, AE Studio July 2023

- Received a grant from AE Studio, a brain-machine interface developing company, to write report analyzing different potential learning frameworks for learning normative knowledge/human values (report found above in "Project" section.)

References

Francesca Zaffora Blando (fzaffora@andrew.cmu.edu)
Carnegie Mellon University

Aybüke Özgün (a.ozgun@uva.nl)
University of Amsterdam/ILLC

Tom Schoonen (t.schoonen@uva.nl)
University of Amsterdam/ILLC

Daniel Herrmann (d.a.herrmann@rug.nl)
University of Groningen

Richard Korzekwa (rick@aiimpacts.org)
Machine Intelligence Research Institute/AI Impacts