

Joe Carlsmith

joe@joecarlsmith.com – www.joecarlsmith.com

EDUCATION

Oxford University (St. John's College)

- DPhil in Philosophy, Fall 2019 – Spring 2023.
- Dissertation: “A Stranger Priority: Topics at the Outer Reaches of Effective Altruism.” Link [here](#). Supervised by Prof. Hilary Greaves and Prof. Jeff McMahan.

New York University

- PhD student in Philosophy, Fall 2016 – Spring 2018 (I spent the academic year of 2017-18 in Oxford).

Oxford University (Merton College)

- BPhil/MPhil in Philosophy, June 2014. *Distinction*.
- Thesis Title: “Hypocrisy and Accountability,” supervised by Prof. John Gardner.

Yale University

- Bachelor of Arts, Philosophy (with honors), May 2012. *Summa cum laude*.

EMPLOYMENT

- *Anthropic*, Member of Technical Staff, Fall 2025-present, working on the design of the production model's constitution.
- *Visiting Lecturer*, Yale Law School, spring term of 2026, co-teaching a course on “Regulating deceptive and scheming AI.”
- *Coefficient Giving* (then: Open Philanthropy), Fall 2018-Fall 2025. Senior Advisor (2025); Senior Research Analyst (2021-2025); Research Analyst (2018-2021).
- Research Assistant to Dr. Toby Ord for his book [The Precipice: Existential Risk and the Future of Humanity](#), *Future of Humanity Institute/Centre for Effective Altruism*, Fall 2017-Summer 2018.

FELLOWSHIPS/AWARDS

- Future of Humanity Institute DPhil Scholarship – University of Oxford
- MacCracken Fellowship – New York University
- Clarendon Scholarship – University of Oxford
- Warren Memorial High Scholarship Prize, awarded to the Yale senior majoring in the humanities who ranks the highest in scholarship.

PUBLICATIONS

- “Claude’s Constitution,” by Amanda Askell, Joe Carlsmith, Chris Olah, Jared Kaplan, Holden Karnofsky, several Claude models, and many other contributors, *Anthropic*, January 2026.
- “Existential Risk from Power-Seeking AI,” in *Essays on Longtermism*, eds. Hilary Greaves, Jacob Barrett, and David Thorstad, Oxford University Press, 2025.
- “Scheming AIs: Will AIs fake alignment during training in order to get power?”, *Open Philanthropy Report*, published on arXiv Nov 2023, <https://arxiv.org/pdf/2311.08379>.

- “Is Power-Seeking AI an Existential Risk?”, *Open Philanthropy Report*, April 2021, published on arXiv June 2022, <https://arxiv.org/abs/2206.13353>. Video summary [here](#). Slides [here](#). Reviews [here](#).
- “How Much Computational Power Does It Take to Match the Human Brain?”, *Open Philanthropy Report*, September 2020, link [here](#), blog post summary [here](#).

OTHER WRITING

- *How do we solve the alignment problem?*, essay series (in progress), 2025-2026.
- *Otherness and control in the age of AGI*, book-length essay series, 2024.
- Papers from the dissertation:
 - “[SIA vs. SSA](#)”
 - “[Simulation arguments](#)”
 - “[Infinite ethics and the utilitarian dream](#)”
- Example other essays:
 - “[The stakes of AI moral status](#)”
 - “[Predictable updating about AI risk](#)”
 - “[Can you control the past?](#)”
 - “[Why should ethical anti-realists do ethics?](#)”
 - “[Seeing more whole](#)”
 - “[On sincerity](#)”
 - “[On expected utility](#)”
 - “[Against neutrality about creating happy lives](#)”
- Full list of essays [here](#).

PRESENTATIONS

- “How human-like do safe AI motivations need to be?”
 - *Constellation*, December 2025 (video and transcript [here](#)).
 - *Far AI alignment workshop*, December 2025.
- “Deflationary vs. validating theories of consciousness, and their implications,” *Eleos AI welfare Conference*, November 2025.
- “Degrees of Agency,” panel discussion at *The Curve* with Ketan Ramakrishnan and Joshua Rothman, Berkeley, October 2025.
- “Giving AIs safe motivations,” *UT Austin AI and Human Objectives Institute*, September 2025, video and transcript [here](#).
- “Can goodness compete?”
 - Keynote at the *Post-AGI Civilizational Equilibria Workshop*, Vancouver, July 2025, video [here](#).
 - *Public talk at Mox*, San Francisco, July 2025, video and transcript [here](#).
- “How do we solve the alignment problem?”
 - *Google DeepMind*, AGI and Society speaker series, May 2025.
 - *LessOnline*, June 2025.
 - *Manifest*, June 2025.
- “How should we think about AI welfare?”
 - *UT Austin AI and Human Objectives Institute*, September 2025.
 - *Manifest*, June 2025.
 - *Anthropic*, May 2025, video and transcript [here](#).
- “Can we safely automate alignment research?”
 - *Anthropic*, April 2025, video and transcript [here](#).
 - *Constellation*, April 2025.

- “Consciousness, robust agency, or something else?,” *AI Welfare Workshop (organized by Eleos AI)*, February 2025.
- “Paths to solving the alignment problem”
 - *Redwood Research Workshop on AI Futurism and Safety*, January 2025.
 - *Constellation AI Safety Organizers Workshop*, January 2025.
- “Otherness and control in the age of AGI,” *Stanford University*, October 2024, video and transcript [here](#).
- “Scheming AIs: Will AIs fake alignment during training in order to get power?” *EA Global 2024*, video and transcript [here](#).
- “Grand Unified Cost-Effectiveness Models,” *Global Priorities Institute Retreat*, July 2022.
- “AI Timelines and Deep Learning,” *Center for Human-Compatible Artificial Intelligence 2022 Conference*, panel discussion with Stuart Russell, Owain Evans, and Sam Bowman, June 2022.
- “Existential Risk and Animal Welfare,” *Research Seminar hosted by LSE’s philosophy department for students*, discussion with Jonathan Birch, November 2021.
- “Is Power-Seeking AI an Existential Risk?” (variants of this talk)
 - *ML Safety Scholars Program*, July 2022.
 - *Harvard Agathon Lecture Series*, March 2022, video [here](#).
 - *Global Priorities Institute*, May 2021.
- “Thoughts on Global Catastrophic Risk Research.”
 - *Swiss Existential Risk Initiative (CHERI)*, August 2021.
 - *Stanford Existential Risk Initiative*, June 2020.
- “Orienting Towards the Long-Term Future.” *Effective Altruism Global* (London), November 2017. Video [here](#).

PODCASTS

- “[A wiser, AI-powered civilization is the ‘successor’](#),” *The Trajectory with Dan Fagella*, November 2025.
- “[Consciousness and competition](#),” on *Forecast with Fin Moorhouse*, October 2025.
- “[Otherness and control in the age of AGI](#),” *The Dwarakesh Podcast*, August 2024.
- “[Joe Carlsmith on Scheming AI](#),” *Hear This Idea with Fin Moorhouse*, March 2024.
- “[On Infinite Ethics, Utopia, and AI](#),” *The Foresight Institute Existential Hope Podcast*, September 2023.
- “[The Dangers of Advanced AI: An Existential Risk?](#)” *The Flares Podcast*, July 2023.
- “[Navigating Serious Philosophical Confusions](#),” *The 80,000 Hours Podcast with Rob Wiblin*, May 2023.
- “[How We Change Our Minds About AI Risk](#),” *Future of Life Institute Podcast with Gus Docker*, June 2023.
- “[Utopia, AI, and Infinite Ethics](#),” *Lunar Society Podcast with Dwarakesh Patel*, August 2022.
- “[Utopia on Earth and Morality Without Guilt](#),” *Clearer Thinking with Spencer Greenberg*, August 2021.
- “[Creating Utopia](#),” *The Utilitarian Podcast with Gus Docker*, July 2021.

PROFESSIONAL SERVICE

- Co-organizer (with Prof. Jeff McMahan and Ketan Ramakrishnan), *Derek Parfit Memorial Conference*, Oxford University, May 2018.
- Co-organizer (with Prof. Samuel Scheffler), *NYU Ethics Forum* (series of talks in value theory at NYU), Spring 2017.