AI Risks:

From Today to Doomsday

Hosted by the Amsterdam AI Safety Initiative.

Sponsored by ELLIS.

Agenda

- 18:00-18:45 **Keynote by Ajeya Cotra and Q&A**
- 18:45-19:45 **Panel discussion and Q&A**
- After 19:45 **Social at Oerknal**

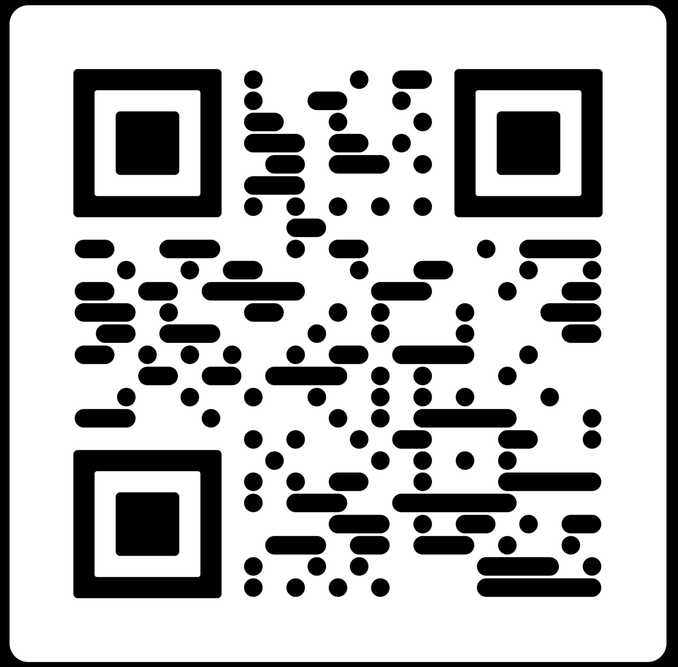
Keynote

Ajeya Cotra

- B.Sc. in Electrical Engineering and Computer Science at UC Berkeley.
- Joined Open Philanthropy as a Research Analyst in 2016.
- Now leads Open Phil's grantmaking on technical research for reducing catastrophic risks from advanced AI.
- Analyzes threat models and technical agendas for advanced AI.
- Estimated transformative AI timelines in her well-known biological anchors report.

Audience Q&A

- **Slido:** Join at slido.com with code: 1653725
- Or with QR

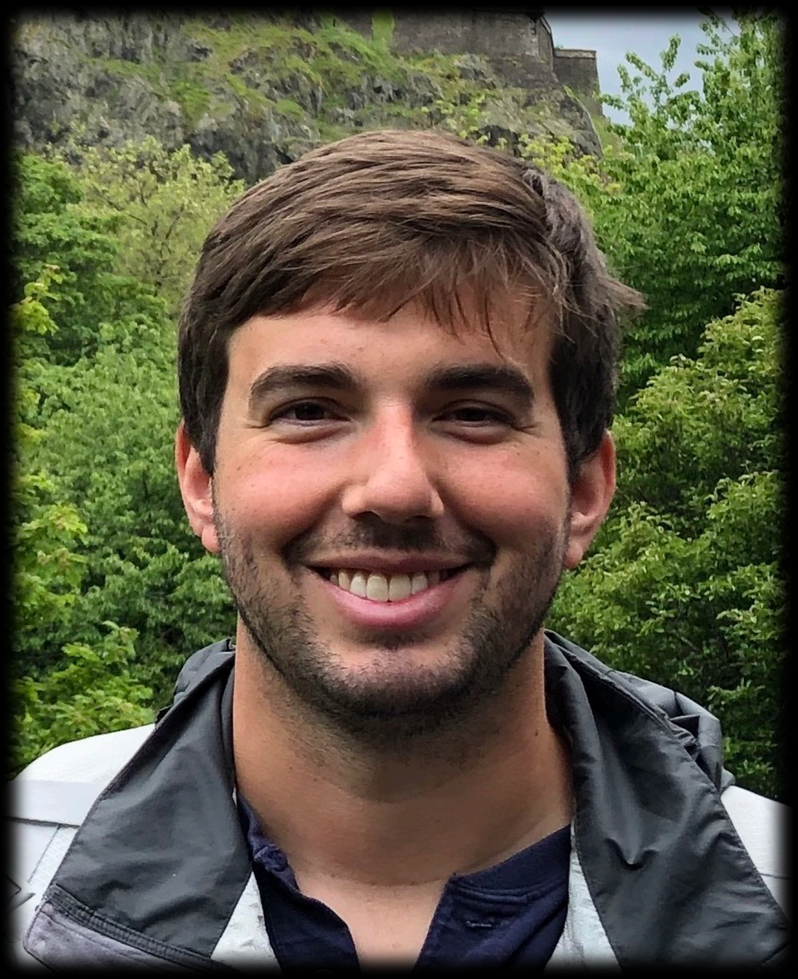


Panel Discussion

Tim Bakker

- PhD Researcher in Machine Learning at UvA.
- Focuses on active learning and active sensing.
- Interests span from Bayesian probability theory to musical theater.
- Recently talked publicly on the existential risks of AI.

Eric Nalisnick



- Assistant Professor at the UvA’s Informatics Institute.
- Focuses on probabilistic modeling, computational statistics, and incorporating human prior knowledge into AI systems.
- Recognized as an ELLIS scholar and an NWO Veni fellow for responsible AI research.
- Former Research Scientist at Google DeepMind and interned at Microsoft, Twitter, and Amazon.

Iris Groen

- Assistant Professor at the UvA’s Informatics Institute.
- Utilizes EEG, fMRI, and ECoG to study human visual perception and computational models.
- Completed postdocs at New York University and the National Institutes of Mental Health.
- MacGillavry Fellowship recipient, funded by Veni Grants and interdisciplinary PhD programme grant from the UvA's Data Science Center.

Jakub Tomczak

- Associate Professor at TU/e, previously at Vrije Universiteit Amsterdam and the UvA.
- Previously at Qualcomm AI Research.
- Deep generative modeling and Bayesian inference.
- Author of the book *Deep Generative Modeling* and founder of Amsterdam AI Solutions.

Panel Discussion

I. The Risk Landscape

II. Responding to the Landscape

III. Looking Forward

I. The Risks Landscape

- **Understanding the Risks:** Of the many potential AI risks — from misuse, centralisation of power, and the propagation of societal biases, to automated capitalism and AI takeovers — which scenarios do you find most worrying and why?
- **Grasping Misalignment**: AI alignment often deals with models not meeting the intentions of their designers. Can you break down the challenge of AI alignment? How should we perceive scenarios of misalignment and the spectrum of potential outcomes?

II. Responding to the Landscape

- **Navigating AI's Dual Nature:** How do you personally strike a balance in this double-edged research space, and are there specific areas you prioritize or actions you take?
- **Technical Solutions and Broader Strategies:** Are there any technical solutions you believe are particularly promising? Are there aspects of the problem that cannot be solved by technical solutions?

III. Looking Forward

- **Humanity's Trajectory:** How do you envision the trajectory of humanity in a world where machines surpass us in overall capabilities?
- **Conceptualizing Transformative AI:** How do you conceptualize transformative AI, and when do you anticipate we might reach that threshold? Do you think current models and techniques are steering us in the right direction?

Thank you!

Join Us: AI Safety Initiative Amsterdam

https://aisafetyamsterdam.com/

- **WhatsApp** Announcements and Community Chats.
- **AI Safety Fundamentals** course twice a year.
- More to come…