

ETH ZURICH

27 SEPTEMBER, 2025

# AI SAFETY DAY

CONFERENCE  
PROGRAM

ORGANIZED BY

# AGENDA

09:00 – 09:50 Doors open + registration

10:00 – 10:35 **Welcome + Opening keynote**

10:35 – 10:55 Coffee Break + Room Transitions

10:55 – 11:20 **Innovation vs. Safety: A False Dichotomy**

*Anastasiia Gaidashenko (FAR.AI)*

11:25 – 11:45 **Getting Started in AI Safety: A Guide for Newcomers**

*Artem Petrov (Palisade)*

12:00 – 14:00 **Org fair + Lunch**

14:00 – 15:15 **Parallel Tracks I**

Scan for the  
live program



## Technical AI Safety

14:00 – 14:20 | **Hidden Failures: When Models Deceive Their Evaluators**

*Johannes Gasteiger (Anthropic)*

14:25 – 14:45 | **Training OpenAI's frontier models not to scheme**

*Teun van der Weij (Apollo Research)*

14:50 – 15:15 | **Lighting talks by UK AISI**

*Art O Cathain, Lennart Luetzgau & Magda Dubois (UK AISI)*

## Governance

14:00 - 15:15 | **The Geopolitics of AGI**

*Alix Pham, Daan Juijn (Simon Institute & centre for future generations)*

## Careers

14:00 - 15:15 | **Challenges in Mid-Career Transitions & What to Expect from Career Advising.**

*Moneer Moukaddem (Successif)*

15:15 – 15:30 **Break**

15:30 – 16:45 **Parallel Tracks II**

## Technical AI Safety

15:30 – 16:45 | **AI Evaluation Discussion Panel with Apollo, Palisade, & UK AISI**

*Justin Olive, Artem Petrov, Alex Lloyd*

## Governance

15:30 – 16:00 | **Fireside chat: From Zurich to Brussels and back - lessons from the EU policy bubble.**

*Tekla Emborg, Santeri Koivula (FLI)*

## Careers I

15:30 – 16:45 | **Pathways to Impact: What Really Gets Professionals into AI Safety Careers.**

*Nina Friedrich (HIP)*

## Careers II

15:30 – 16:45 | **Round table with a career advisor: Personalized career transition advice.**

*Moneer Moukaddem (Successif)*

17:00 – 17:45 **Closing session**

17:45 - 18:45 **Aperó + Networking**