AISI AI SECURITY INSTITUTE | ANTHROP\C | APOLLO RESEARCH | FAR.AI | Palisade Research

ETH ZURICH

27 SEPTEMBER, 2025

AI SAFETY DAY

CONFERENCE PROGRAM

ORGANIZED BY

BlueDot Impact

ZURICH AI SAFETY

AGENDA

Scan for the live program

09:00 – 09:50 Doors open + registration | HG Vorhalle

10:00 – 10:20 **Opening session** | HG F1

10:20 – 10:45 **Innovation vs. Safety: A False Dichotomy**
Anastasiia Gaidashenko (FAR.AI) | HG F1

10:45 – 10:55 Coffee Break + Room Transitions | HG EO Nord

10:55 – 11:20 **Democratizing Open and Compliant LLMs For Global Language Environments**
Imanol Schlag (Apertus, ETH AI Center) | HG F1

11:25 – 11:45 **Getting Started in AI Safety: A Guide for Newcomers**
Artem Petrov (Palisade) | HG F1

12:00 – 14:00 **Org fair + Lunch** | CLA Glashalle + Clausiusbar

14:00 – 15:15 ***Parallel Tracks I***

Technical AI Safety

14:00 – 14:20 | **Hidden Failures: When Models Deceive Their Evaluators**
Johannes Gasteiger (Anthropic) | HG F1

14:25 – 14:45 | **Training OpenAI's frontier models not to scheme**
Teun van der Weij (Apollo Research) | HG F1

14:50 – 15:15 | **Lighting talks by UK AISI**
Art O Cathain, Lennart Luettgau & Magda Dubois (UK AISI) | HG F1

Governance

14:00 – 14:45 | **Building a technical roadmap for the effective governance of AI**
Nandini Shiralkar (ERA Fellowship) | HG E1.1

Careers

14:00 - 15:15 | **Challenges in Mid-Career Transitions & What to Expect from Career Advising**. *Moneer Moukaddem (Successif)* | HG E1.2

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**
Alex Lloyd, Justin Olive, Artem Petrov, & Magda Dubois | HG F1

Careers I

15:30 – 16:45 | **Pathways to Impact: What Really Gets Professionals into AI Safety Careers.** *Nina Friedrich (HIP)* | HG E1.2

Careers II

15:30 – 16:45 | **Round table with a career advisor: Personalized career transition advice.** *Moneer Moukaddem (Successif)* | HG E22

17:00 – 17:45 **Closing session + Closing Keynote** | ML D28
Adam Gleave (FAR.AI)

17:45 - 18:45 **Aperó + Networking** | CLA Glashalle