

RESEARCH HACKATHON 2026

SCIENTIFIC RESEARCH INSTEAD OF MVPs

RESEARCH FRAME

“Loss of control through autonomous agents”

DATES

20-22 NOVEMBER 2026

VENUE

FÜRTH · NUREMBERG REGION

PARTICIPANTS

UP TO 100 RESEARCHERS & STUDENTS

TEAMS

3-5 PEOPLE OR SOLO SIGN-UP

01 EXECUTIVE SUMMARY & VISION

COAI deliberately breaks with the product-driven hackathon format. Instead of immature prototypes or MVPs, the focus is on substantial, open research questions.

The three-day sprint aims at robust scientific starting points: precise hypotheses, sound methodology, well-founded arguments or first experimental data. Promising approaches are to continue in a structured mentoring programme towards publication (workshop or conference papers).

TECH & POLICY ON EQUAL FOOTING

Code is not a prerequisite. Rigorous governance, legal and policy analyses (e.g. regulatory scenarios, gap analyses) have the same methodological standing as technical implementations.

02 THE RESEARCH FRAME

No isolated tracks in 2026: every team works under one shared, highly topical frame, viewed through three established perspectives.

DETECT

Early-warning systems, metrics and anomaly detection for drift or goal deviation in agentic systems.

UNDERSTAND

Systemic root-cause analysis, interaction patterns and false assumptions in the oversight of agents.

CONTROL

Technical and regulatory safeguards: guardrails, emergency shutdown, intervention architectures.

03 TOOLING & FORMATS

AGENT-COLLABS SANDBOX · OPTIONAL

A modular multi-agent test environment for experimental work: one isolated instance per team, with example scenarios such as a trojan or sleeper within a collective.

RECOMMENDED OUTPUT

1

STRUCTURED ABSTRACT

Max. 600 words in six sections, from scenario and thesis via method and result to outlook and references; appendix for code or logs.

2

ADEQUATE EVIDENCE

Run logs, code repositories, source collections or coherent legal-conceptual argumentation.

3

PITCH

6 slides · 5-minute talk · 3-minute expert Q&A.

OPEN RESEARCH · NOT JUDGED

OPEN STAGE

A jury-free presentation stage for innovative work outside the official research frame.

04 INTERDISCIPLINARY JURY & EVALUATION RUBRIC

The jury spans technology, policy, mechanistic interpretability and AI safety. Scientific quality is assessed with one rubric:

CRITERION	WEIGHT	CORE QUESTION
Contribution to the theme	25 %	Is a real mechanism of loss of control addressed, with a robust gain in insight?
Method & reasoning	25 %	Is the approach adequate and transparently documented, with its limitations reflected?
Evidence	20 %	Are the claims backed by valid data, runs, cases or sources?
Scientific novelty	15 %	Does the work go significantly beyond known demos and standard reviews?
Science communication	15 %	Are question and results conveyed clearly and precisely across disciplines?