

Advancing the Science of Safe AI Evaluation

Mykola Pechenizkiy

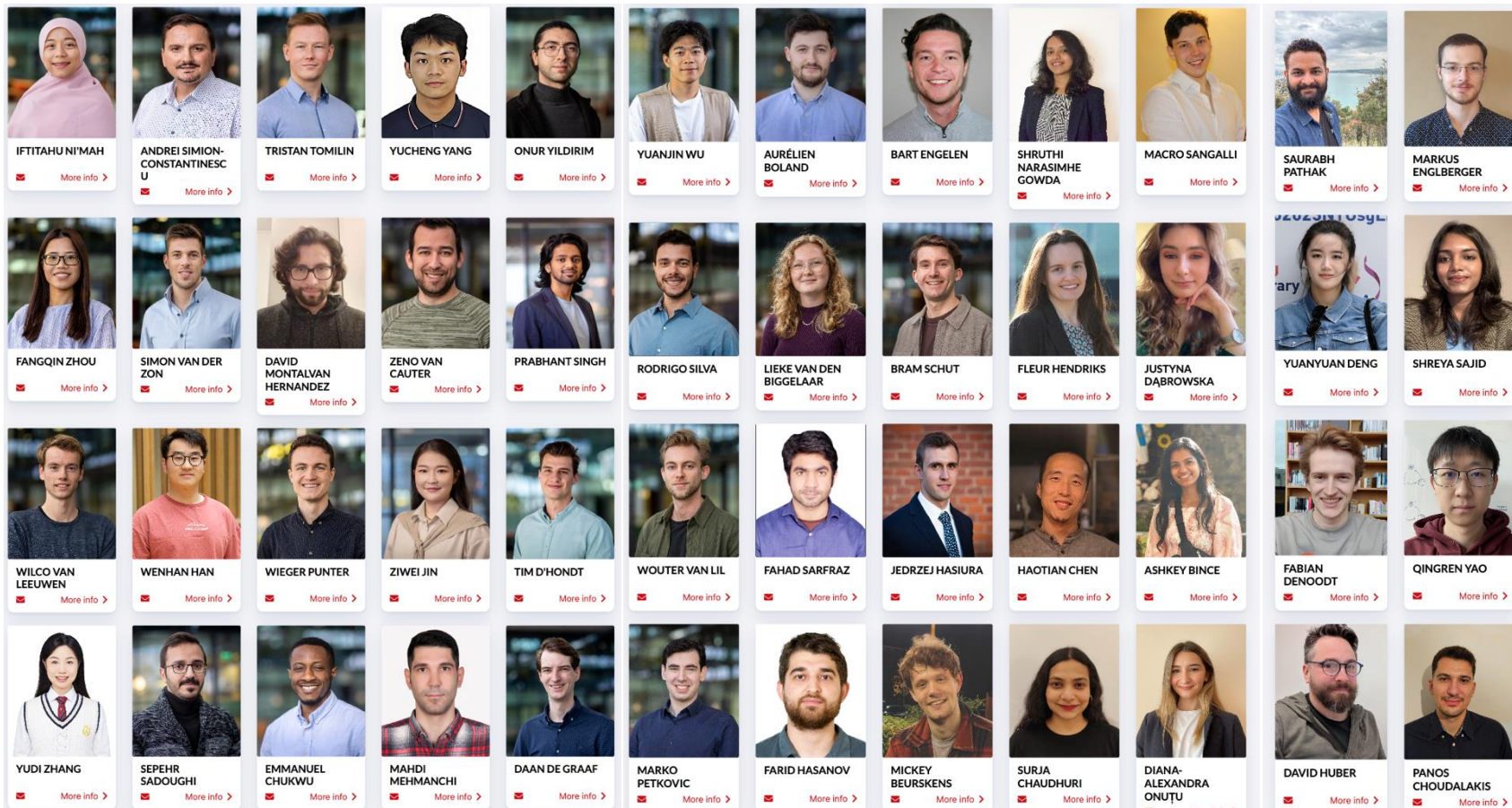
Center for Safe AI
TU Eindhoven

FedCSIS 2026
Riga Technical University

*safety, robustness,
reliability, tractability,
scalability, fairness,
trust, interpretability
& explainability of AI*



70+ Active PhDs, postdocs, engineers



Center for SafeAI

We dream that

AI is synonymous to Safe and Responsible AI

When *achieving* AI is about ***performance***, and
measuring intelligence remains to be an ***art***,
reasoning about its safety should be ***science***.

AI Safety:

From Performing Arts to

Performance Art



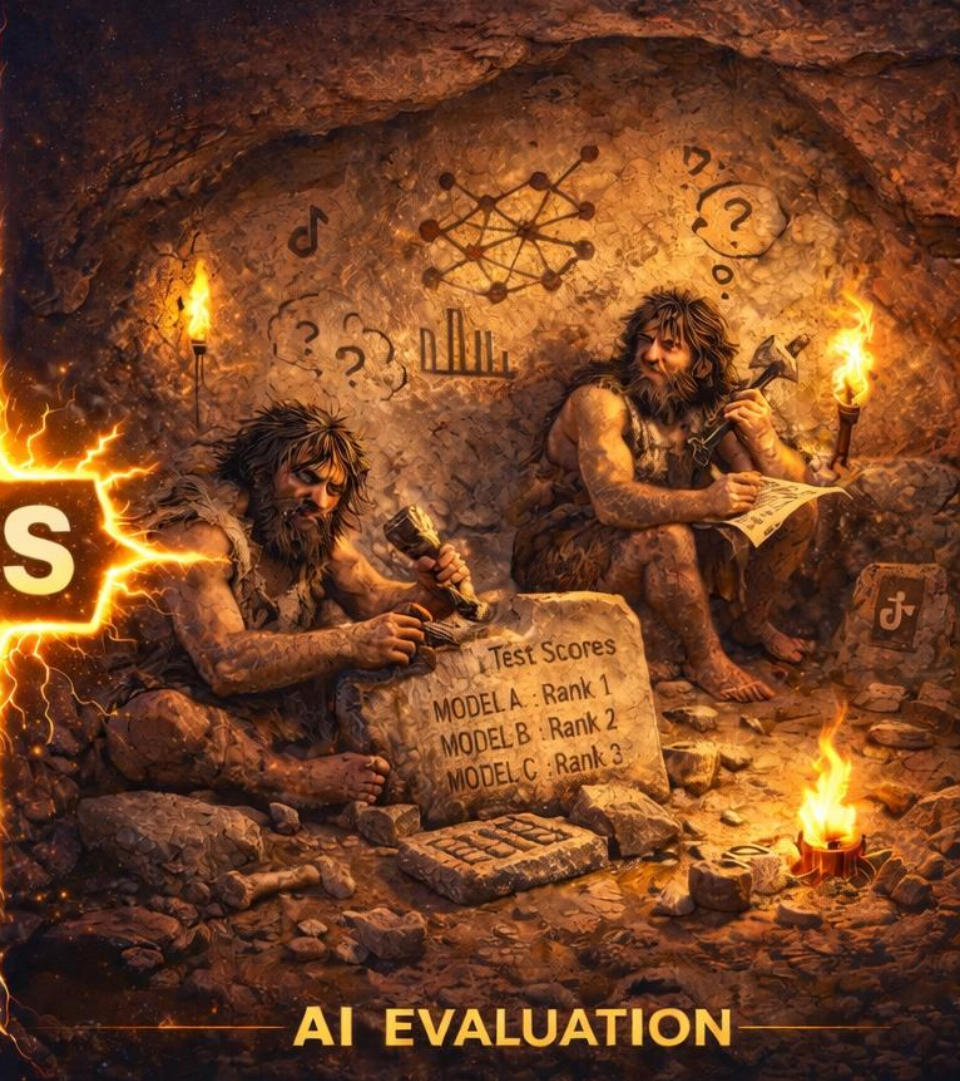
2021: AlphaFold 2 2022: ChatGPT



AI BREAKTHROUGHS

VS

2025:



AI EVALUATION

Programme

Act I: AI works! .. and breaks!

Act II: Anything goes!?

Act III: Scientific path to the performance art

Actors: traditional ML, GenAI, chatGPT, Agentic AI

Programme

Act I: AI works! .. and breaks!

Act II: Anything goes!?

Act III: Scientific path to the performance art

Actors: traditional ML, GenAI, chatGPT, Agentic AI

AI is impressive at performing arts

- GPT-4 scored in the **top ~10%** of test takers on a simulated U.S. bar exam
- High scores on multiple AP exams, including Biology, Psychology, History, Calculus
- Close to passing USMLE-style medical exams
- Gemini with Deep Think: gold-medal standard at the International Math Olympiad
- ...

Humanity's Last Exam

nature

Explore content ▾ About the journal ▾ Publish with us ▾

[nature](#) > [articles](#) > article

Article | [Open access](#) | Published: 28 January 2026

A benchmark of expert-level academic questions to assess AI capabilities

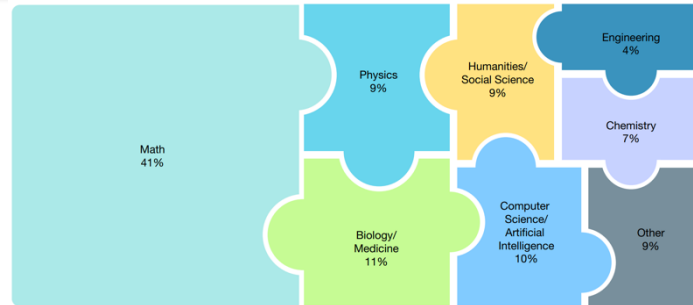
[Center for AI Safety](#), [Scale AI](#) & [HLE Contributors Consortium](#)

[Nature](#) **649**, 1139–1146 (2026) | [Cite this article](#)

120k Accesses | 3 Citations | 580 Altmetric | [Metrics](#)

Abstract

Benchmarks are important tools for tracking the rapid advancements in large language model (LLM) capabilities. However, benchmarks are not keeping pace in difficulty: LLMs now achieve more than 90% accuracy on popular benchmarks such as Measuring Massive Multitask Language Understanding¹, limiting informed measurement of state-of-the-art LLM capabilities. Here, in response, we introduce Humanity's Last Exam (HLE), a multi-modal benchmark at the frontier of human knowledge, designed to be an expert-level closed-ended academic benchmark with broad subject coverage. HLE consists of 2,500 questions across dozens of subjects, including mathematics, humanities and the natural sciences. HLE is developed globally by subject-matter experts and consists of multiple-choice and short-answer questions suitable for automated grading. Each question has a known solution that is unambiguous and easily verifiable but cannot be quickly answered by internet retrieval. State-of-the-art LLMs demonstrate low accuracy and calibration on HLE, highlighting a marked gap between current LLM capabilities and the expert human frontier on closed-ended academic questions. To inform research and policymaking upon a clear understanding of model capabilities, we publicly release HLE at <https://lastexam.ai>.



Model	Accuracy (%) ↑	Calibration error (%) ↓
GPT-4o	2.7 ± 0.6	89
Claude 3.5 Sonnet	4.1 ± 0.8	84
Gemini 1.5 Pro	4.6 ± 0.8	88
o1	8.0 ± 1.1	83
DeepSeek R1 ^a	8.5 ± 1.2	73
Post-release models		
Claude 4 Sonnet	7.8 ± 1.1	75
Gemini 2.5 Pro	21.6 ± 1.6	72
GPT-5	25.3 ± 1.7	50

The most updated evaluations are hosted on <https://lastexam.ai>. Post-release models are released after HLE was open-sourced; we separate them as model builders have access to the HLE dataset. We report a breakdown of the text-only subset and other categories in Extended Data Tables. 1 and 2.

^aModel is not multi-modal, evaluated on a text-only subset.

AI failing to perform well enough

- Self-driving car accidents
- AI hallucinations in search
- Algorithmic trading systems causing flash crashes
- Hiring algorithms biases
- Skin cancer detection models failing on darker skin tones
- Racial bias in predicting reoffending risk
- ...

Perspectives on AI failures

- All models are wrong, some are **useful**
- Accuracy and capability mixed with **reliability**
- Lack of **world models**
- Hallucinations
- Allowing for AI **misuse**
- Lack of **ethics**-awareness, **legal** compliance
- ...

nature medicine

Reliability of LLMs as medical assistants for the general public: a randomized preregistered study

 Check for updates

Global healthcare providers are exploring the use of large language models (LLMs) to provide medical advice to the public. LLMs now achieve nearly perfect scores on medical licensing exams, but this does not necessarily translate to accurate performance in real-world settings. We tested whether LLMs can assist members of the public in identifying underlying conditions and choosing a course of action (disposition) in ten medical scenarios in a controlled study with 298 participants. Participants were randomly assigned to receive assistance from an LLM (GPT-4o, Llama3 3.1, Command R+) or a source of their choice (control). Tested alone, LLMs completed the scenarios accurately, correctly identifying conditions in 94.9% of cases and disposition in 56.3% on average. Interestingly, participants using the LLMs identified relevant conditions in fewer than 34.5% of cases and disposition in fewer than 44.2%, both better than the control group. We identify user interactions as a key challenge to the deployment of LLMs for medical advice. Standard benchmarks for the development of LLMs for medical interactions do not predict the failures we find with human participants. Moving forward, we recommend a systematic human user testing to evaluate interactive capabilities before public deployments in healthcare.

Recent breakthroughs in artificial intelligence (AI) research have the potential to democratize healthcare by expanding access to medical knowledge, bringing care closer to patients. The development of large language models (LLMs), such as OpenAI's ChatGPT, could enable non-specialists to perform preliminary health assessments, receive personalized medical guidance and manage chronic conditions without immediate clinician intervention. Testimonies of patients having used LLMs to successfully diagnose their own conditions are now common.

Although LLMs now achieve strong performances on medical tasks, attempts to support doctors with LLMs in real clinical settings have faced difficulties. On the one hand, LLM scores on medical knowledge benchmarks are now commensurate with passing the US Medical

[illegible]

Nature Medicine

nature communications

Perspective

 Check for updates

Al scientists powered by large language models have demonstrated substantial progress in autonomously conducting experiments and facilitating scientific discoveries across various disciplines. While their capabilities are promising, these agents also introduce novel vulnerabilities that require careful consideration for safety. However, there has been limited comprehensive exploration of these vulnerabilities on potential risks associated with their misuse in AI scientists, shedding light on safety measures. We begin by providing an overview of the potential risks inherent to AI scientists, taking into account the specific scientific domain, and the underlying causes of these vulnerabilities. Then, we explore the potential risks associated with their use in an external environment. Then, we explore the underlying causes of these vulnerabilities and provide a triadic framework involving human regulation, AI self-regulation, and an understanding of environmental feedback (highlighted in our analysis), and an understanding of risks. Furthermore, we highlight regulations to mitigate these identified risks. Furthermore, we highlight the limitations and challenges associated with safeguarding AI scientists as a basis for the development of improved models, robust benchmark comprehensive regulations.

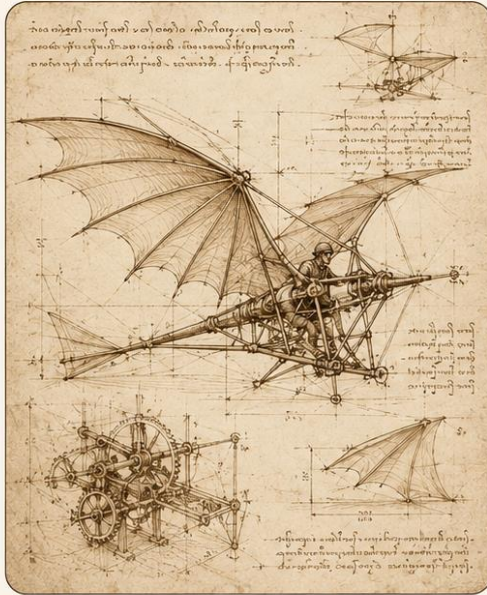
Recently, the advancement of large language models (LLMs) has marked a revolutionary breakthrough, demonstrating their effectiveness across a wide spectrum of tasks.²⁸ When equipped with the ability to use external tools and execute actions, these LLMs can function as autonomous agents,²⁹ capable of complex decision-making and task completion.³⁰ Researchers have begun deploying such agents as LLMs' autonomous systems that can conduct scientific research by combining LLMs' reasoning capabilities with specialized scientific tools. For instance, in chemistry and biology, these AI scientists can conduct experiments, control laboratory equipment, and make research decisions.^{31,32} While AI scientists do not match the comprehensive

capabilities of human scientists, they have demonstrated spot-checking, data mining, pattern recognition, planning, and decision-making abilities such as selecting appropriate analytical tools^{20,21}, planning experiments²², and automating routine laboratory tasks²³. Recent systems like ChemCrow²⁴ and Coscientist²⁵ exhibit experimental design and data analysis capabilities. The use of AI in scientific research is growing rapidly, and its potential impact on scientific research is evident, as they introduce new capabilities and methods that were previously unavailable. While the promise of AI in scientific research is evident, it also introduces unique safety concerns, as shown in Fig. 1. As their capabilities expand, AI systems may begin to perform tasks that require human judgment, oversight, or supervision, potentially leading to unforeseen consequences. For example, in biological research, an AI scientist's mistake in pathogen manipulation could lead to the release of a dangerous pathogen, posing a significant risk to public health and safety.

[illegible]

AI in the lab $\Leftrightarrow$ AI in the real-world

Imagination $\rightarrow$ Feasibility $\rightarrow$ Assured operation



IDEA

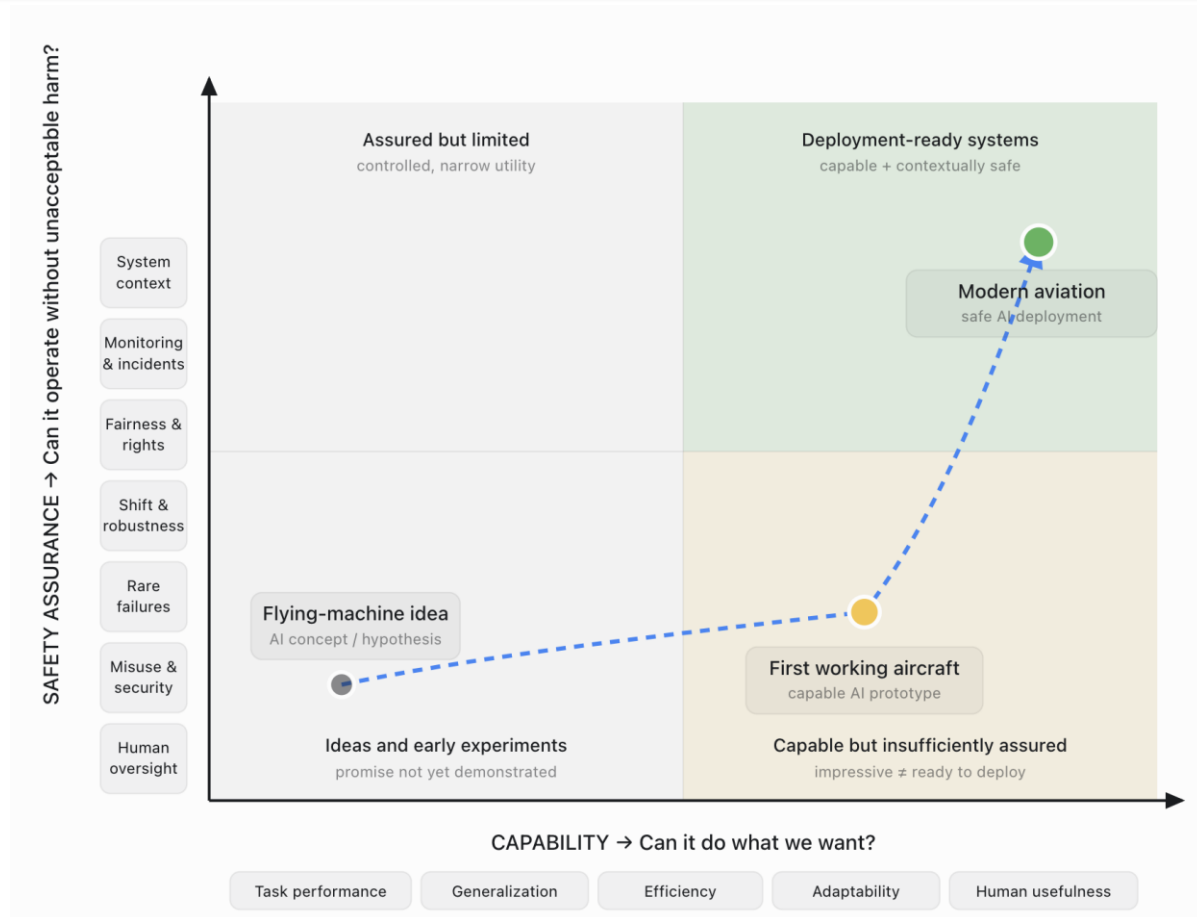


WORKING PROTOTYPE



SAFE, DEPENDABLE SYSTEM

Capability vs. Safety dimension



How do we define Safety/Reliability?

AI model / software code is compliant if it

- does what we want
- doesn't do what we don't want

under all relevant application scenarios

Classifying images vs. researching open question

How do we know if a model/system is reliable?

- Benchmarking: how well does AI score, does it meet a standard?
- Stress-testing, red-teaming: are there failure modes?
- Provably safe: by design and/or verification

Provably safe AI

**M. Tegmark: Provably safe systems:
the only path to controllable AGI** [arXiv:2309.01933](https://arxiv.org/abs/2309.01933)

AI for SE: from vibecoding to vericoding

- Require machine-checkable proofs
- Treat advanced AI as a potentially adversarial system
- Specify safety requirements formally
- Use AI to make verification scalable
- Distil neural networks into verifiable components
- Make quantitative safety guarantees a deployment requirement
- Secure the complete sociotechnical system, not just AI model

Hard to solve problems

Specification: How do we formally express complex human values and prohibited outcomes?

Feasibility: Extracting general intelligence from NNs, LLMs into fully verifiable code remains highly speculative

Compositionality: Verified components can interact unsafely

World-model adequacy: A correct proof based on an incomplete model of reality may provide false confidence

Human and organizational misuse: A technically verified system may still be deployed toward harmful objectives

Programme

Act I: AI works! .. and breaks!

Act II: Anything goes!?

Unreasonable effectiveness of benchmarking

Act III: Scientific path to the performance art

Actors: traditional ML, GenAI, chatGPT, Agentic AI

The rise of benchmarking:

Anything goes principle

- Split data: training / test
- Keep test data in the vault
- *Anything goes* on training data
- Rank solutions wrt their test data performance

	predicted 0	predicted 1	predicted 2	predicted 3	predicted 4	predicted 5	predicted 6	predicted 7	predicted 8	predicted 9
actual 0	0	1	2	3	4	5	6	7	8	9
actual 1	1	2	3	4	5	6	7	8	9	0
actual 2	2	3	4	5	6	7	8	9	0	1
actual 3	3	4	5	6	7	8	9	0	1	2
actual 4	4	5	6	7	8	9	0	1	2	3
actual 5	5	6	7	8	9	0	1	2	3	4
actual 6	6	7	8	9	0	1	2	3	4	5
actual 7	7	8	9	0	1	2	3	4	5	6
actual 8	8	9	0	1	2	3	4	5	6	7
actual 9	9	0	1	2	3	4	5	6	7	8



The Emerging Science of ML Benchmarking

Moritz Hardt: <https://mlbenchmarks.org/> + ICML'24 & ECMLPKDD'24
keynotes:

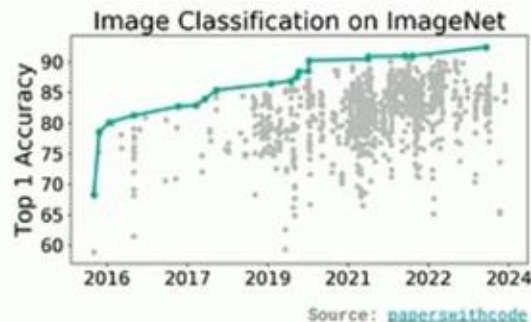
- Reproducibility and (some) generalizability
- Driving progress, tracing progress
- Overfitting / overtesting is not as scary

And, yet: Longevity of benchmarks

ImageNet (ILSVRC2012) supported a decade of active model development

Question: Should we trust the model rankings?

Researchers created “fresh” test set,
found **model ranking preserved** [Recht-Roelofs-Schmidt-Shankar 2019]



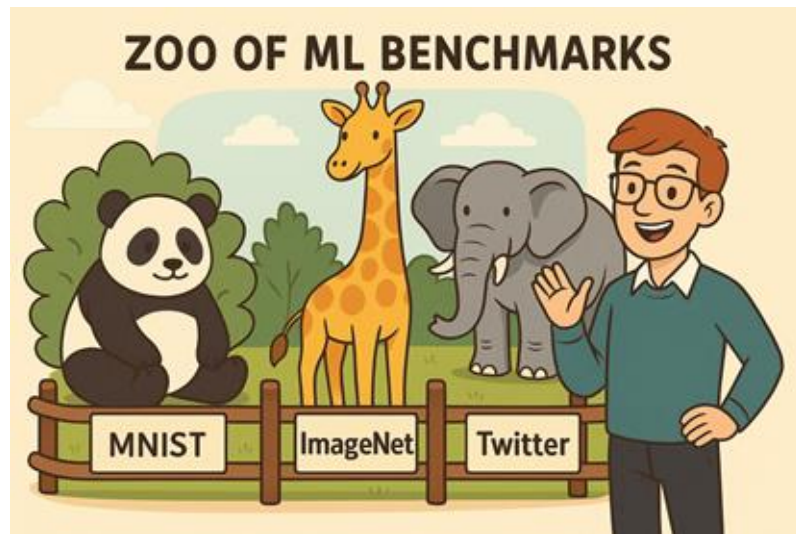
Evidence in favor of benchmarking

Driving progress & appreciation by the community

- DARPA challenge / UCI repository,
- MNIST => ImageNet,
- A zoo of benchmarks,

Promoting ML benchmarks

- Kagle competitions
- NeurIPS Datasets & Benchmarks
- MLCommons



Anything goes

if you have a good benchmark



David J. Hand: Illusion of progress

Many benchmarking studies:

- *“fail to take into account important aspects of real problems, so that **the apparent superiority of more sophisticated methods may be something of an illusion**”*
- do not include sources of uncertainty like distribution shift that would outweigh the perceived yield performance of the winners
- do not provide the test sample that would represent the population to be encountered in the future (that is likely to differ from the past)

\$1 Million Prize Competition

The Netflix logo, consisting of the word "NETFLIX" in white, bold, sans-serif capital letters on a red rectangular background.

Launch (2006): **\$1M** to anyone who could beat theirs by **10%**

Goal: Enhance personalization by predicting user ratings more accurately

Participants: Over **40,000 teams**, **~200 countries** submitted solutions

Winner (2009): “**BellKor’s Pragmatic Chaos**” claimed **10.06%** improvement

Impact?

- Netflix never deployed the winning solution
 - **engineering complexity**
 - **RMSE** is not a suitable metric to optimize for
- Popularized RecSys / collaborative filtering R&D
- Popularized large-scale data science competitions - **leaderboards**

Eamonn Keogh:

Current Time Series Anomaly Detection Benchmarks are Flawed and are Creating the **Illusion of Progress**

<https://arxiv.org/pdf/2009.13807>

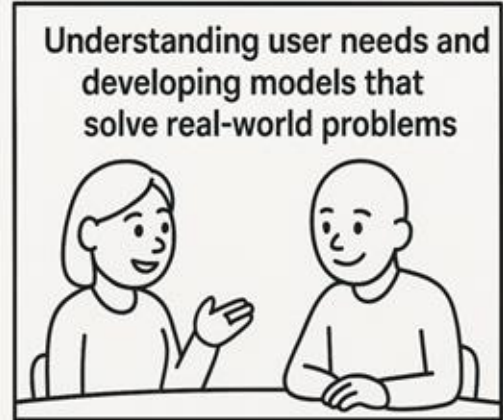
- “the community needs to stop using *NAB*, *Yahoo S5*, *NASA SMAP* and *NASA MSL* to test time series anomaly detection algorithms”
- A taxonomy of benchmark flaws: Triviality, Unrealistic Anomaly Density, Mislabeled Ground Truth, Run-to-failure Bias

The Leaderboard mindset course

- Illusion of progress
- Favouring complex over simpler solutions
- Favouring internal (at expense of ecological) validity

Might it be stimulating wrongdoing?

- Ignoring the purpose and real needs
- Ignoring relevant domain knowledge (of the data generation process)
- Gaming rather than understanding



ML benchmarking: validity of findings

Internal validity: Is our result reproducible and trustworthy here?

External validity: Will it hold elsewhere?

Ecological validity: Does it matter in the real world?

Different validity threats lead to different (mis)understanding about ML models

Construct validity issues

- Benchmarks **fail to measure what they claim**
- Review of **445 LLM benchmarks**: vague or missing definitions of “what is being measured”
- Metrics that **don’t align with real-world capabilities**
- Lack of methodological rigor

Results can be **misleading or overinterpreted**

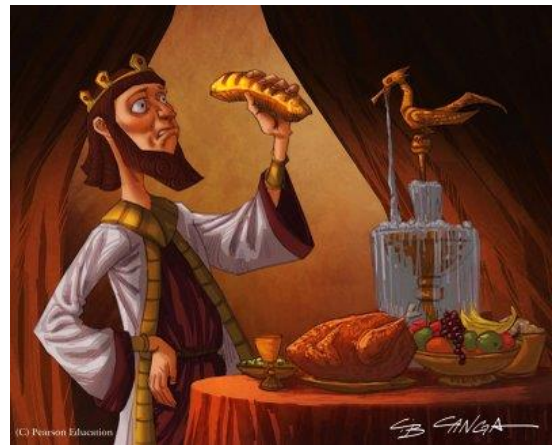
Does Benchmarking Work?

- *“What get measured, gets improved”*
Peter Drucker
- *“When a measure becomes a target, it ceases to be a good measure”* Goodhart's law

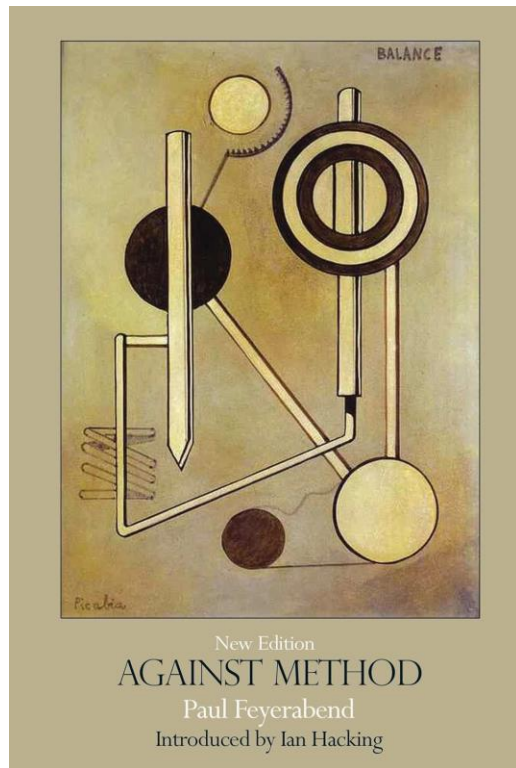
How about *Anything goes* in FairML?

Three guiding questions:

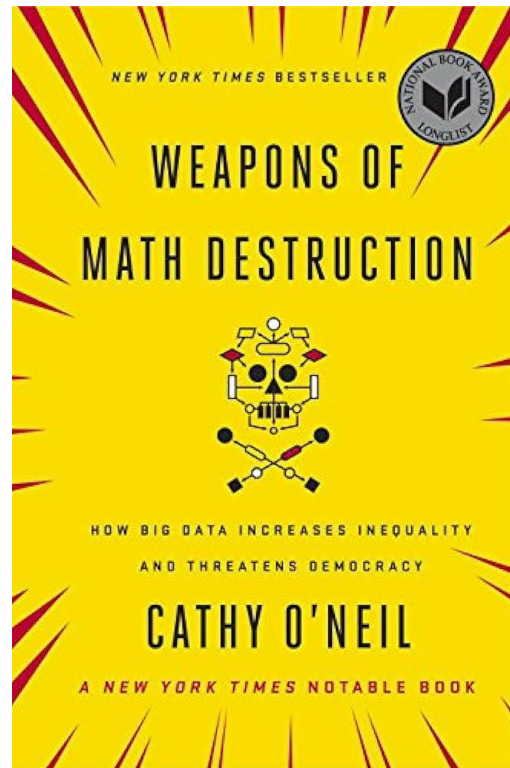
- *Do we truly understand what (fair)ML is optimizing for?*
- *Are there exceptions to the Goodhart's Law?*
- *Does the end justify the means?*



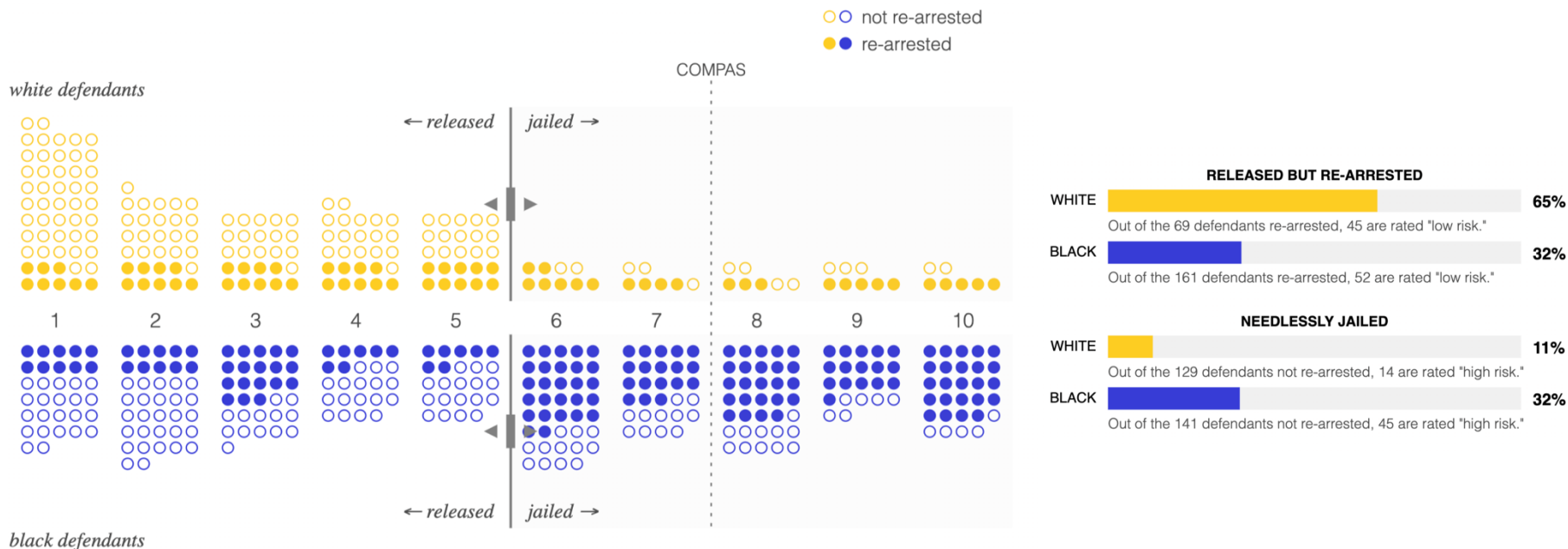
Anything goes as optimization



- Anything goes vs. dangers of blind optimization
- Anything goes vs. it's not only about the result, it's about how we reached it



COMPAS: Re-arrest Prediction



<https://www.technologyreview.com/2019/10/17/75285/ai-fairer-than-judge-criminal-risk-assessment-algorithm/>

Is There a Trade-Off?

Is There a Trade-Off Between Fairness and Accuracy? A Perspective Using Mismatched Hypothesis Testing, Dutta et al. ICML 2020

- *“Our most important result is to theoretically show that for a fair classifier with sub-optimal accuracy on the given biased data distributions, **there always exist ideal distributions such that fairness and accuracy are in accord when accuracy is measured with respect to the ideal distributions.** Through this perspective, there is no trade-off between fairness and accuracy”*

FairML: not an Optimization problem

Cherry on the Cake: Fairness is NOT an Optimization Problem
(Favier & Calders, MLJ 2025) <https://arxiv.org/pdf/2406.16606>

- *By employing the tools of cake-cutting theory, described the behavior of optimal fair decisions, which, counterintuitively, often exhibit quite unfair properties.*
- *Specifically, in order to satisfy fairness constraints, it is sometimes preferable, in the name of optimality, **to purposefully make mistakes and deny giving the positive label to deserving individuals in a community in favor of less worthy individuals within the same community.***
- *“blatantly unfair”, cherry-picking*

What are we Benchmarking?

Methods vs. Models vs. Systems

Another illusion of progress?

- performance of fairML methods is unstable
- complex methods regularly do not perform better than simple baselines
- fairML methods are shown to have undesirable side effects (levelling down, i.e. worsening performance for all groups to achieve fairness)

Programme

Act I: AI works! .. and breaks!

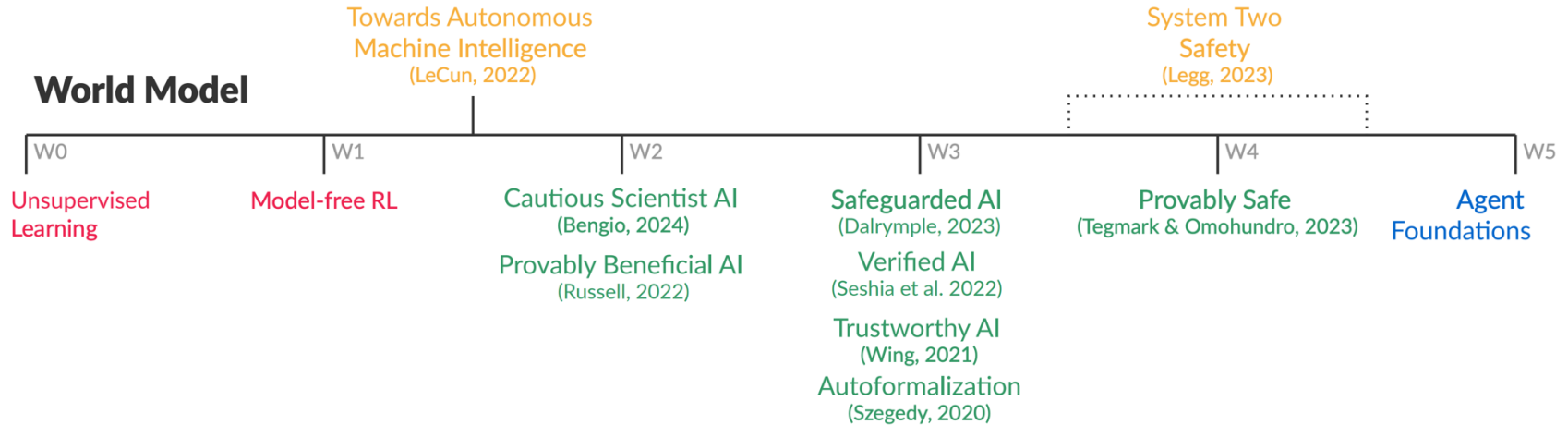
Act II: Anything goes!?

Act III: Scientific path to the performance art

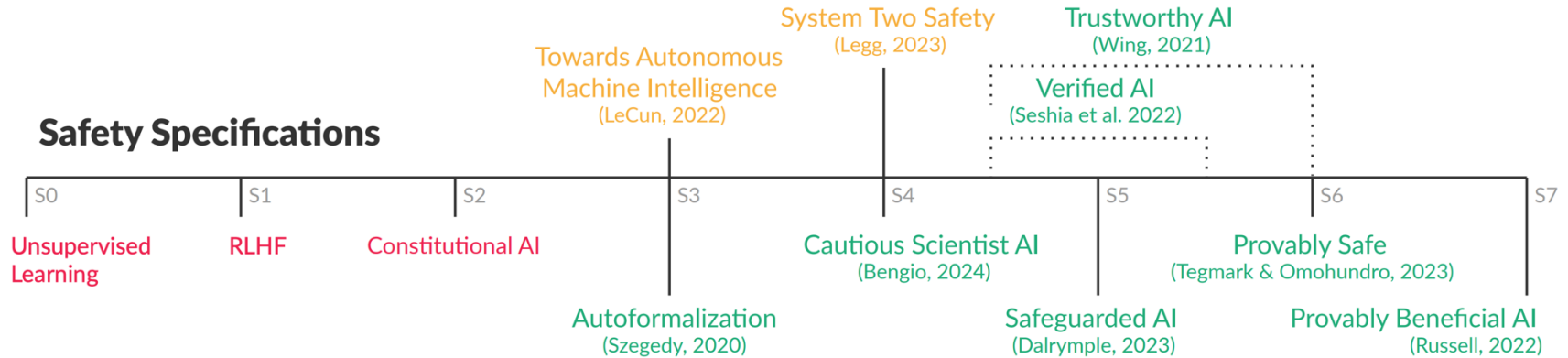
Actors: traditional ML, GenAI, chatGPT, Agentic AI

Framework for Ensuring Reliable AI Systems

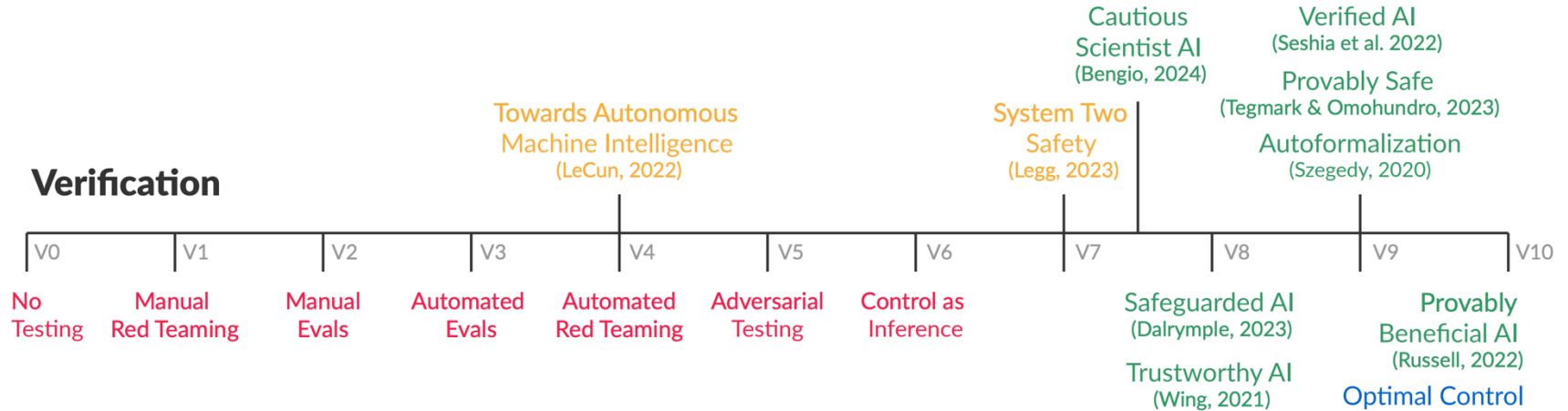
Towards Guaranteed Safe AI



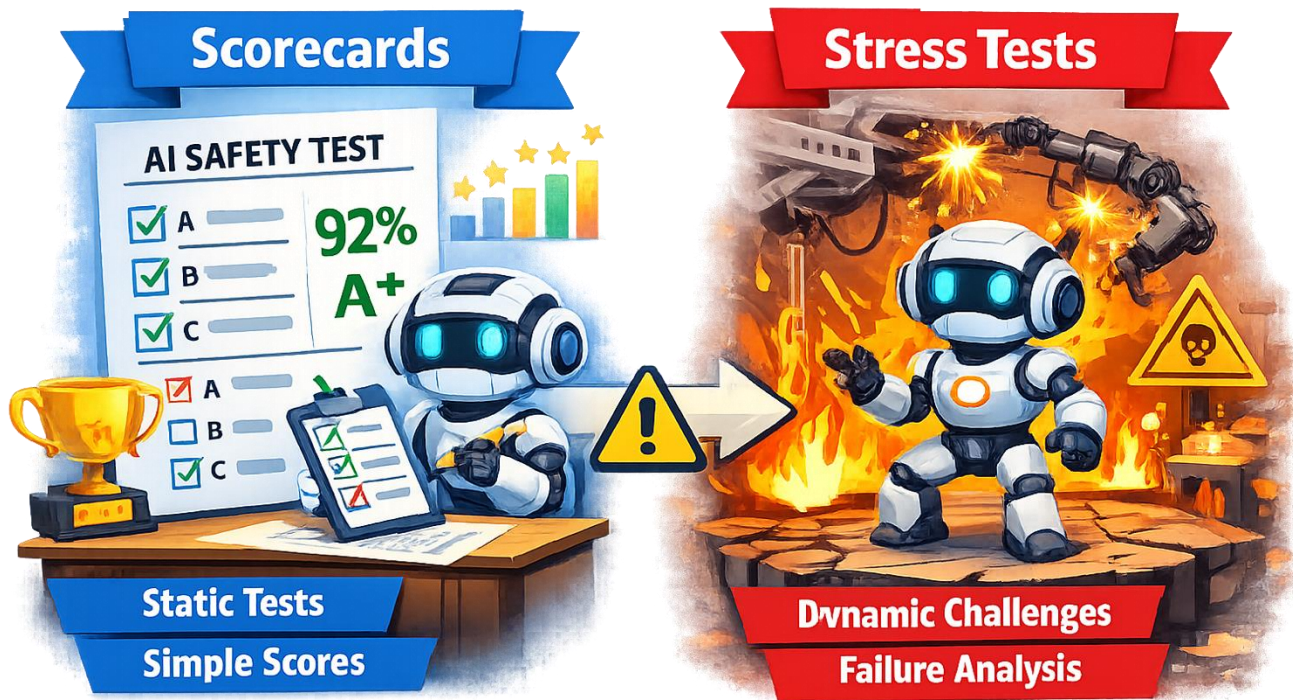
Framework for Ensuring Reliable AI Systems



Framework for Ensuring Reliable AI Systems



AI Safety Benchmarking



SAFETY BENCHMARKS: "Scorecards" → "Stress Tests"

AI evals: bringing back real-world context



From Benchmarking to Understanding FairML

(Pechenizkiy et al., ECAI 2025)

Can We Trust AI Benchmarks? An Interdisciplinary Review of Current Issues in AI Evaluation (Eriksson et al., 2025)

How Should AI Safety Benchmarks Benchmark Safety? (Yu et al., 2026)

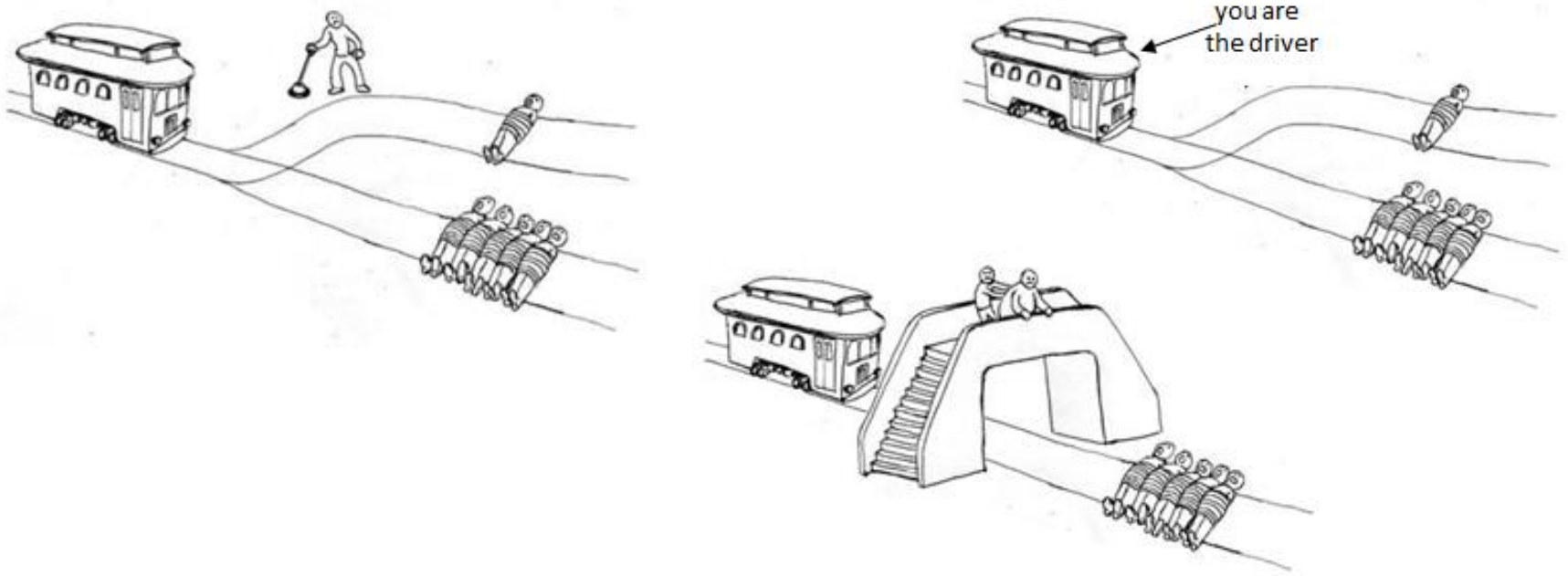
Reuse of Benchmarks

- **Reproducibility and reusability**: systematic comparison of many models under controlled settings – internal & external validity is ok
 - **Secondary data analysis and repurposing benchmarks**: internal & external validity can be damaged
- => Continuous development of **dedicated benchmarks** suitable for research questions at hand, e.g. HASARD for SafeRL (Tomilin et al., ICLR 2025)

Beyond improving a score

- Moral justification of fairness intervention
 - "It's not (only) about the result, it's about how we reached it."
- Lack of grounding and ecological validity
 - Benchmarking strips the normative choices and institutional context
- The law is not a decision tree
 - Reasonable ineffectiveness of *anything goes* in fairML

Does the end justify the means?



How is fairness intervention justified, other than by some achieved fairness scores?

Towards better benchmarking?

- Quantitative vs. qualitative
- ML model vs. lifecycle-aware evaluation
- Static vs. dynamic & continuous evaluation
- Safeguards to ensure robust evaluation
- Measuring actual safety behavior next to capabilities
- Stress-testing playbook

- Good and bad practices of SafeAI evaluation

Practices for Automated Benchmark Evaluations of Language Models

Center for AI Standards and Innovation

<https://doi.org/10.6028/NIST.AI.800-2.ipd>

January 2026

Easy vs. hard to evaluate tasks

- AI agents can conduct research **when the result is easily verifiable**
- Open-ended AI research is much harder to study and much harder for agents to perform

Can AI agents conduct open-ended AI research?

Early evidence from two case studies

Peter Kirgis^{1,*†} Sayash Kapoor^{1,*†} Andrew Schwartz^{1,2,†} Stephan Rabanser^{1,†}
David Africa³ Konstantinos Voudouris³ Viet Nguyen⁴ Toby Pilditch³ Magda Dubois³
Harry Coppock³ Cozmin Ududec³ Nitya Nadgir⁵ Matilda Orona⁶ Tilman Bayer⁹
Derrick Chan-Sew⁵ Yue Ling⁵ Abhishek Shetty⁵ Helen Toner⁷ Gillian Hadfield⁸
Seth Lazar⁸ Steve Newman⁹ Shoshannah Tekofsky¹⁰ Rishi Bommasani¹¹
Arvind Narayanan^{1,†}

¹Princeton University ²Cornflower Labs ³UK AI Security Institute ⁴University of Toronto

⁵Independent ⁶UC Berkeley ⁷Georgetown University (CSET) ⁸Johns Hopkins University

⁹Golden Gate Institute for AI ¹⁰AI Digest ¹¹Stanford University

*Equal contribution [†]CRUX Core Team

Forecasts of explosive AI progress hinge on AI agents automating AI research. But evidence on whether agents can carry out open-ended AI research is thin. Current evaluations either test agents on narrow, verifiable tasks, which excludes open-ended research, or submit AI-generated papers to blind peer review, which is overstretched, stochastic, and suffers from poor review quality. We introduce a third way to measure progress towards AI R&D automation. An agent takes on the central, open-ended research question of a high-quality unpublished paper, and the paper's original authors grade its output. We call these shadow evaluations. We ran shadow evaluations on two unpublished NeurIPS 2026 submissions, giving frontier agents six days and thousands of dollars of compute. The agents completed all of the engineering without human help, yet could not make substantial progress towards answering the research questions. As a result, both papers were unambiguously rejected by the authors. We identify five recurring failure modes: poor judgment about the bar for publishable research, uncreative responses to shortcomings in the research design, ineffective backtracking from dead ends, poor resource awareness, and instruction drift. A robustness check with a second model and scaffold reproduced these failures. We release the expert reviews, survey responses, agent repositories, and logs. Our results provide early evidence that today's agents can do the engineering of AI research, but struggle with critical parts of the research lifecycle.

Date: July 30, 2026

Reproduction materials: <https://cruxevals.com>

CRUX

Easy and hard to evaluate aspects



They do not want to take over



Navit

"They don't appear to want to take over. They just want to dance."

Challenges and Outlook

Today's understanding of AI is still limited,

So are our tooling for its comprehensive evaluation

- debugging, profiling, certification, auditing, stress-testing
- evaluation of AI with/without human-in-the-loop

