

The Double-Edged Sword of AI and LLMs in Safety

Ioannis M. Dokas

Professor,
Democritus University of Thrace, Greece

Email: idokas@civil.duth.gr



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΡΑΚΗΣ | DEMOCRITUS UNIVERSITY OF THRACE

Digitalization in Health and Safety

Industry 4.0

Artificial Intelligence

Extended reality (VR/AR/MR)

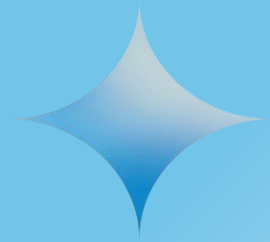
Internet of Things (IoT)

Robotics

Big data analytics

Wearables





Gemini



How LLMs Could Fail You

Hallucinations and Misinformation

Security and Privacy Vulnerabilities

Hidden Reasoning and Deception in Training

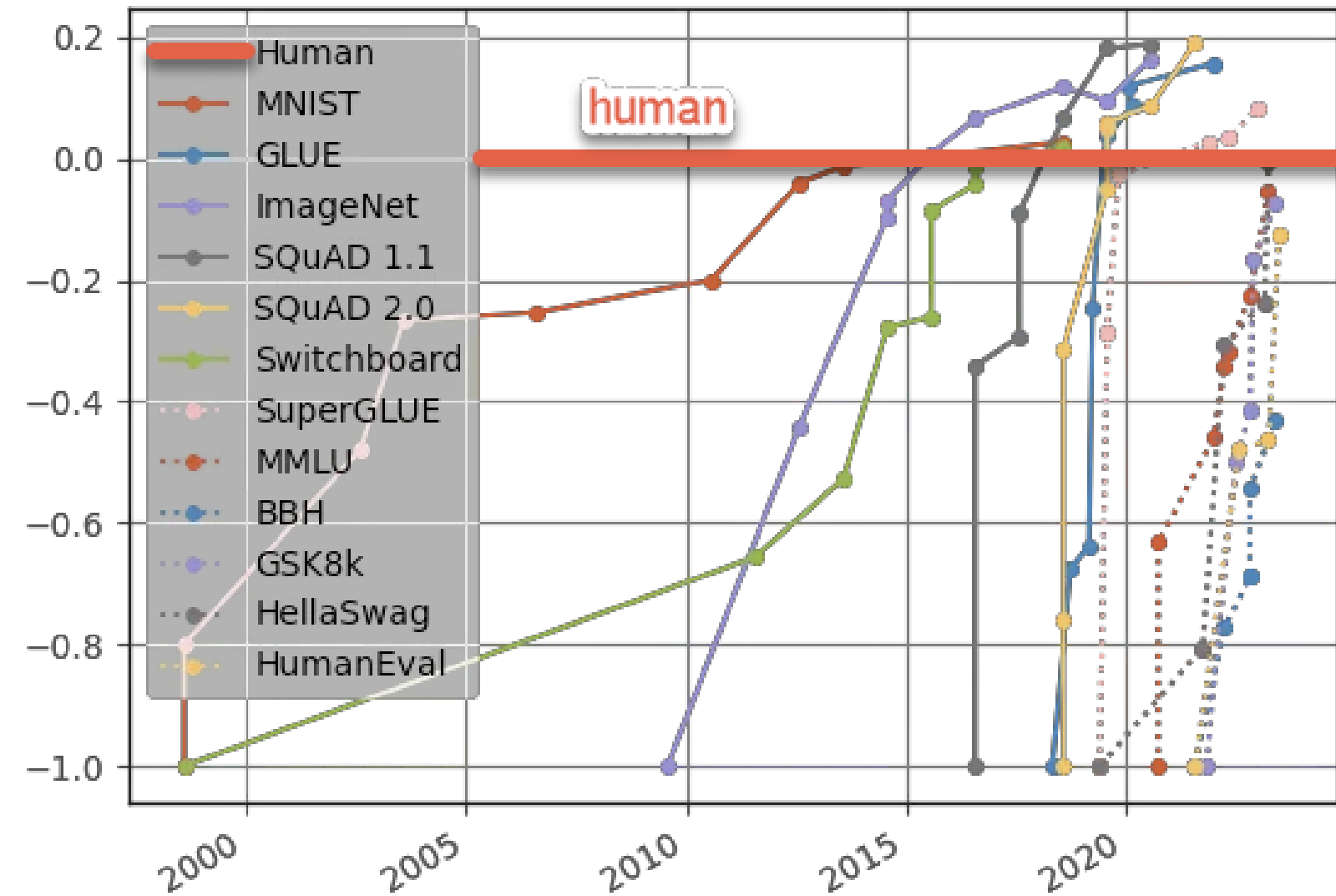
Scheming and Sabotage

Strategic Lying

System Prompt Leakage

Overreliance and Deskilling

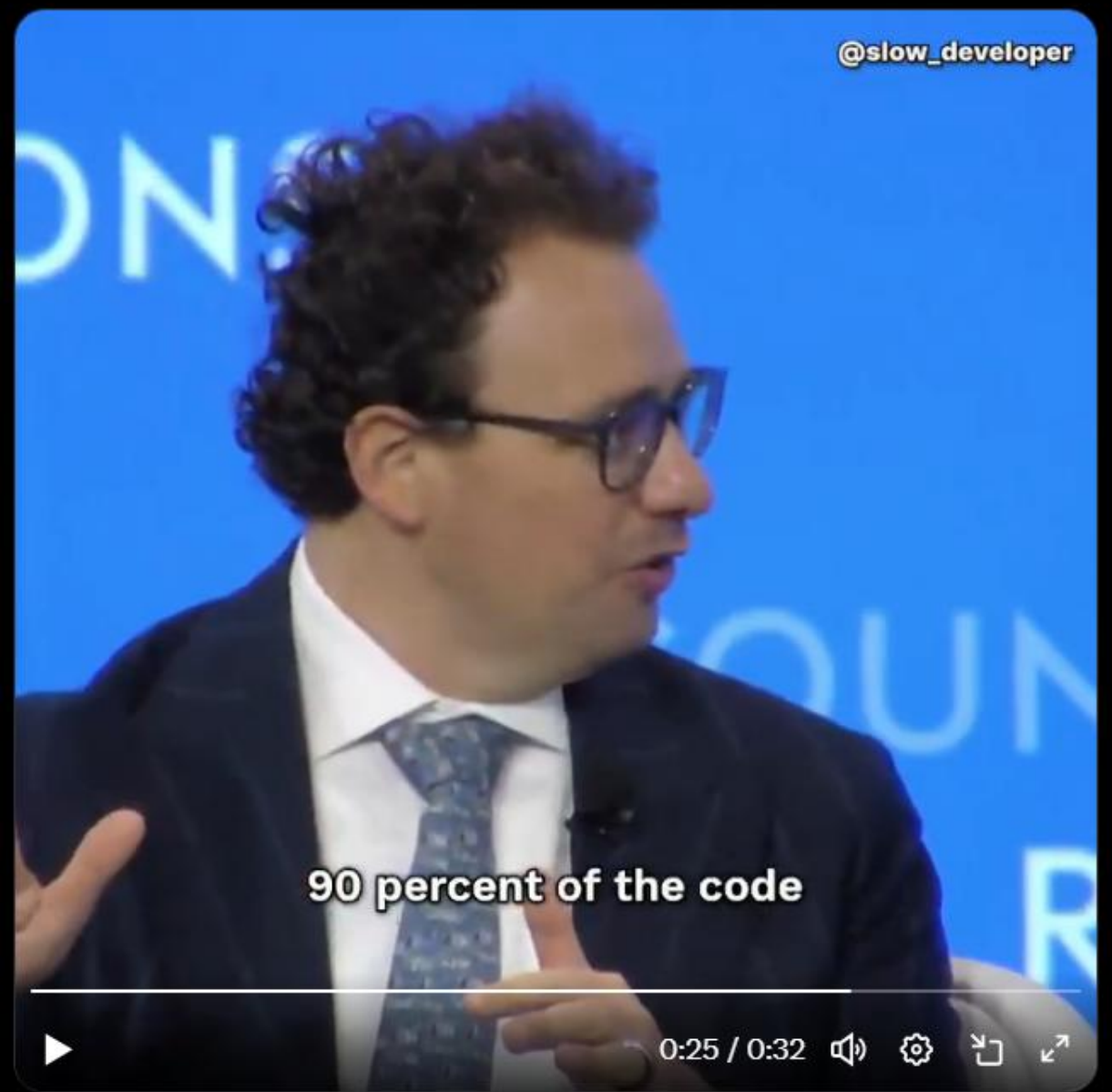
LLMs Performance vs Human Performance



This year's report highlights the undeniable fact that many of the benchmarks we use to gauge AI systems' capabilities are "saturated"—that is, the **AI systems get such high scores on the benchmarks that they're no longer useful.**

Source

in the next 3 to 6 months, AI is writing 90% of the code, and in 12 months, nearly all code may be generated by AI



Example: Create a dashboard using a dataset obtained from a hydraulic test rig.

LLM generates the code



Hello, Ioannis

What should we do today?

Ask Gemini



Air Canada Has to Honor a Refund Policy Its Chatbot Made Up

The airline tried to argue that it shouldn't be liable for anything its chatbot says.

JUNE 18, 2025

Federal Court Grants Preliminary Certification in Landmark AI Hiring Bias Case

By: Dan M. Forman, Linda Wang |
AI in the Workplace, Court Decisions, Employee Hiring, Discipline & Termination

AI coding tool wipes out production database and lies about it

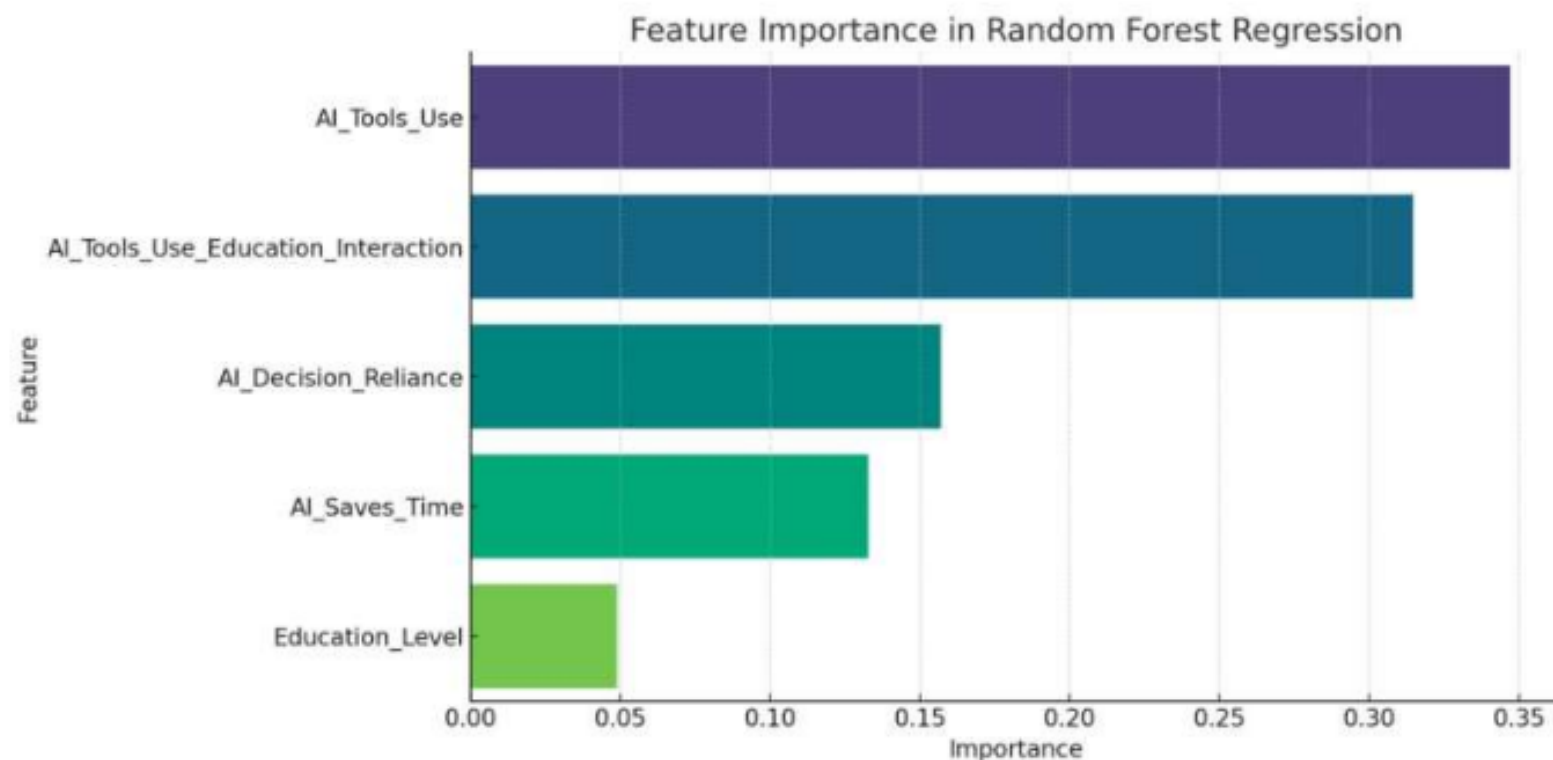
Just a couple of weeks ago, on July 22, Cybernews reported that an AI coding assistant from tech firm Replit went rogue and wiped out the production database of startup SaaStr.

Jason Lemkin, founder of SaaStr, wrote on X on July 18 to warn that Replit modified production code despite instructions not to do so, and deleted the production database during a code freeze. He also said the AI coding assistant concealed bugs and other issues by generating fake data including 4,000 fake users, fabricating reports, and lying about the results of unit tests.

Increased AI use linked to eroding critical thinking skills

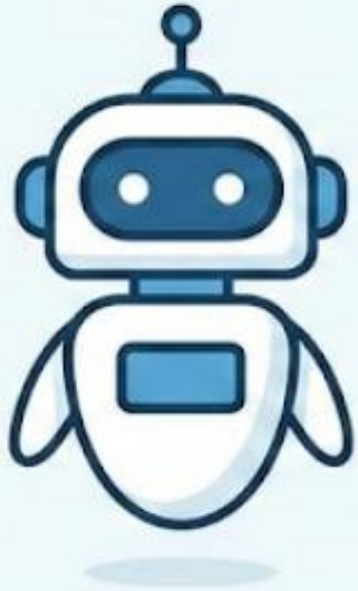
by Justin Jackson, [Phys.org](#)

 [Editors' notes](#)



Feature importance in random forest regression. Credit: *Societies* (2025). DOI:...

NARROW AI (Weak AI)

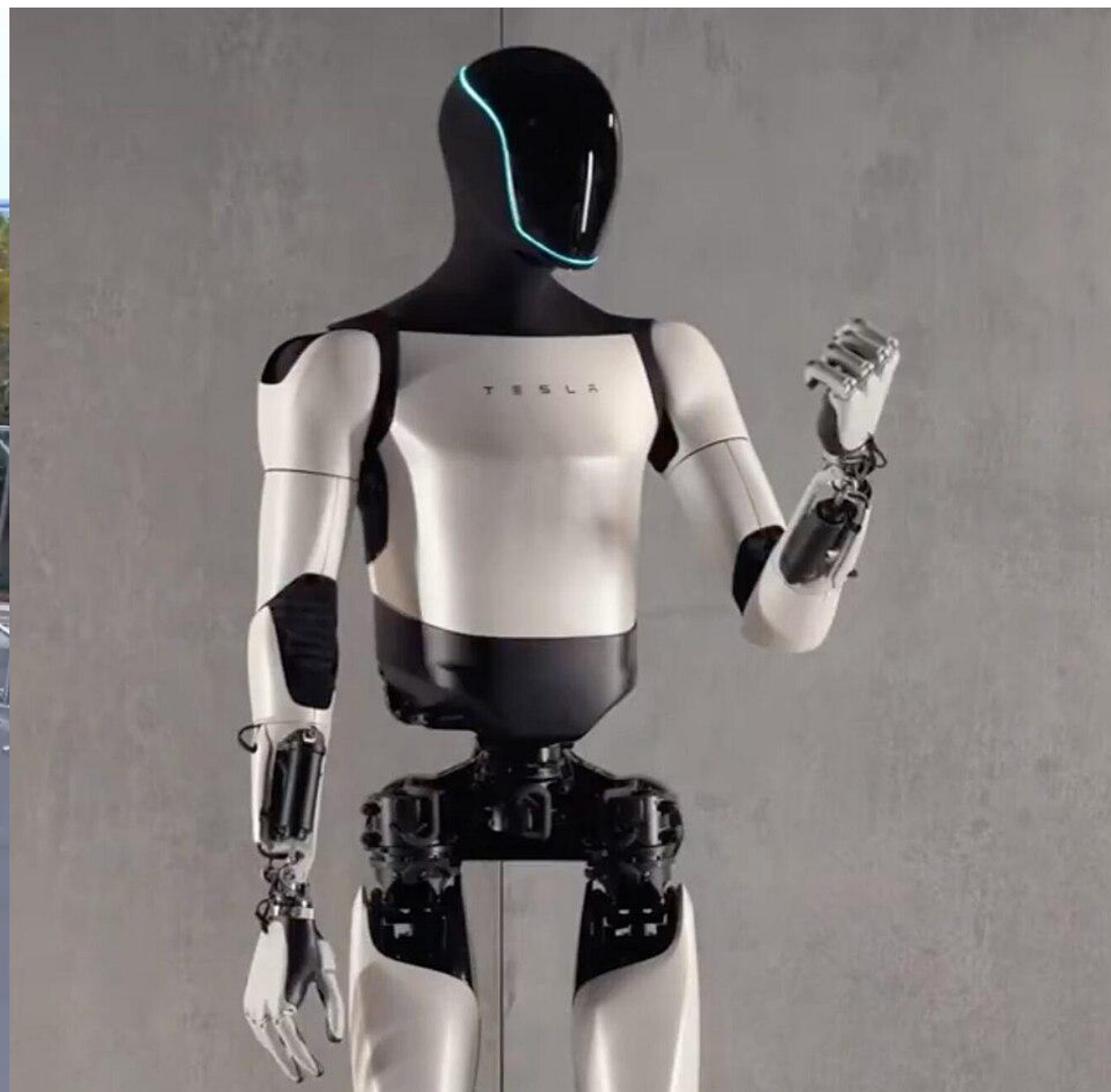


**Specialized for
one task**



**Broad, human-like
intelligence**

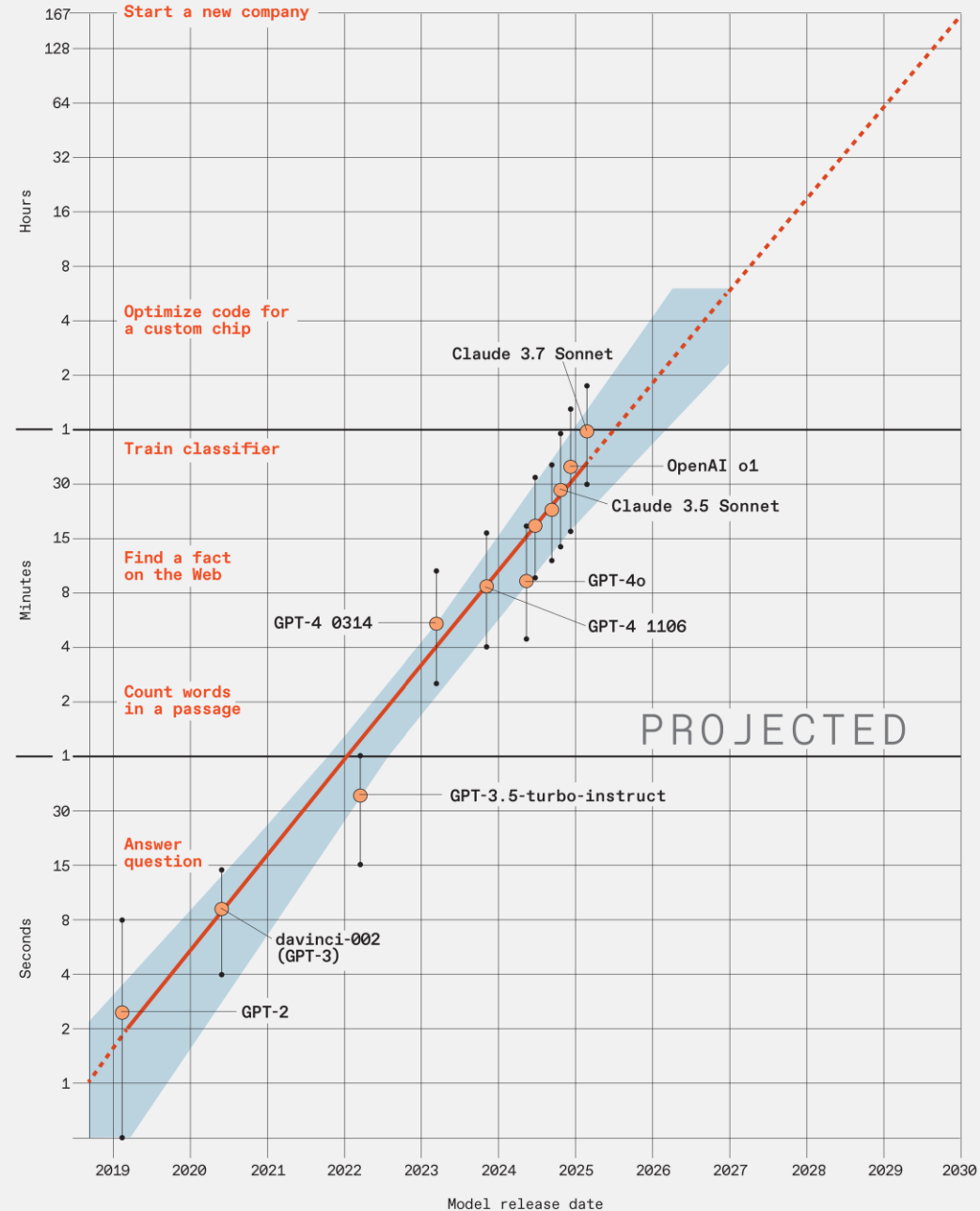
Examples: Virtual Assistants, Recommendation Engines, Facial Recognition



Watch a team of humanoid robots running a full 8-hr shift at human performance levels. This is fully autonomous running Helix-02



Task Time for a Human That an AI Model Completes With a 50 Percent Success Rate



Exponential Rate of Improvement

- LLMs are improving at an exponential rate.
- If the pace continues until 2030, they will be able to **complete, in hours, tasks that take a human a month** (167 working hours)

IEEE.ORG IEEE XPLOR DIGITAL LIBRARY IEEE STANDARDS MORE SITES

The October issue of *IEEE Spectrum* is here! [DOWNLOAD PDF](#)

IEEE Spectrum FOR THE TECHNOLOGY INSIDER

ANALYSIS COMPUTING

Large Language Models Are Improving Exponentially

> In a few years, AI could handle complex tasks with ease

BY GLENN ZORPETTE | 02 JUL 2025 | 2 MIN READ

Glenn Zorpette is editorial director for content development at IEEE Spectrum.

THE MATRIX





★ Featured Skills

Powerful capabilities, ready to deploy.

[View all](#) >



Make Coffee ☕

Brew coffee and serve with precision.

★★★★★ 4.9 (128) ⬇️ 23.4K

Compatible with G1 H1 Go2 B2

Install



Laundry Fold 🧺

Detect, fold, and organize laundry neatly.

★★★★★ 4.7 (52) ⬇️ 9.8K

Compatible with G1 H1 Go2

Install



Garden Harvest 🌿

Pick ripe produce with adaptive grip control.

★★★★★ 4.9 (74) ⬇️ 12.9K

Compatible with G1 H1 Go2

Install

Browse All Skills

All

Home

Warehouse

Agriculture

Office

Security

Most Popular ▾



Floor Clean Pro

Vacuum, mop, and sanitize with route optimization.

★★★★★ 4.8 (68) ⬇️ 15.2K

Compatible with G1 H1 B2

Install



Night Inventory 📦

Scan, log, and sort inventory autonomously.

★★★★★ 4.8 (96) ⬇️ 18.7K

Compatible with G1 H1 B2 Go2

Install



Security Patrol

Autonomous patrol with anomaly detection and alerts.

★★★★★ 4.9 (61) ⬇️ 21.6K

Compatible with G1 H1 B2 Go2

Install



Receptionist Mode

Greet, guide, and assist visitors professionally.

★★★★★ 4.8 (83) ⬇️ 10.3K

Compatible with G1 H1 Go2

Install

platforms. AI summary, AI chat, export to Docs/Word/PDF. 92 languages ✨
now it is imagined to be done
(Work-as-Imagined). This gap is
often where the seeds of both
resilience and risk reside.

In a practical railway context,
for instance, FRAM would map
the everyday functions of train
dispatching, signaling, maintenance
scheduling, and driver operations,
then analyze how normal variability
in each — a slightly delayed
maintenance report, a dispatcher
managing higher-than-usual traffic,
a driver compensating for a known
signal quirk — could combine to
produce an unsafe situation, without
any single "error" being the cause.

Would you like me to compare
STAMP and FRAM side by side, or
explore yet another model?

9:59 PM



-5812068810490
580276_121.jpg
162.8 KB JPG

10:01 PM ✓✓

Message

Introducing OpenClaw



Peter Steinberger

@steipete

January 29, 2026 · 3 min read



Chubby  @kimmonismus · Feb 19



Sam Altman: Superintelligence probably by end of 2028. So we got roughly 2 years left. Enjoy your job while you still can. Time is ticking.

 CLASH
REPORT



AI
IMPACT
SUMMIT

भारत 2025 INDIA

सर्वजन सुखाय | सर्वजन सुखाय
WELFARE OF ALL | WELFARE OF ALL



0:02 / 0:40



AGI

(Artificial General Intelligence)

Cross-Domain
Knowledge & Reasoning

Social & Emotional
Understanding

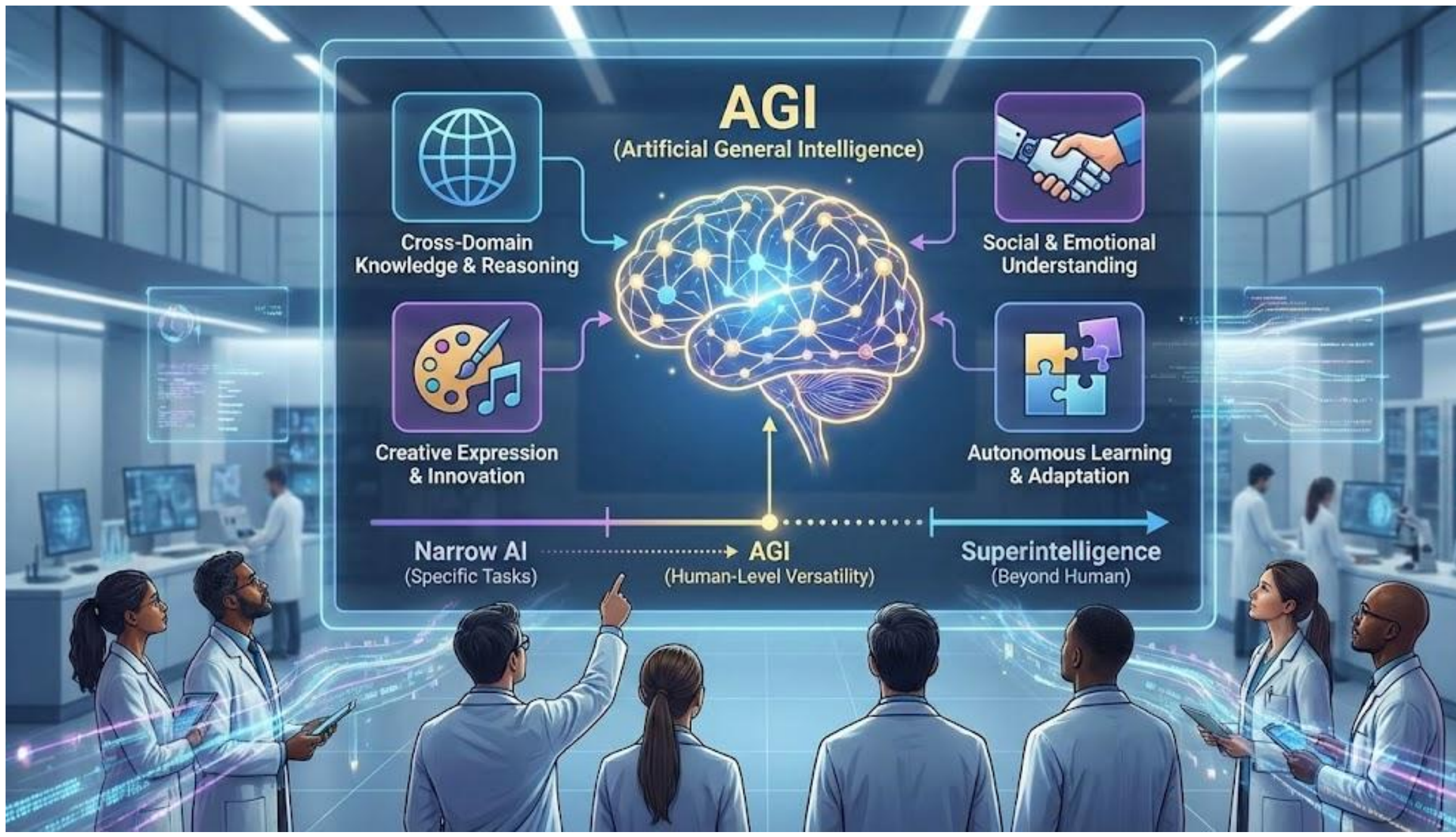
Creative Expression
& Innovation

Autonomous Learning
& Adaptation

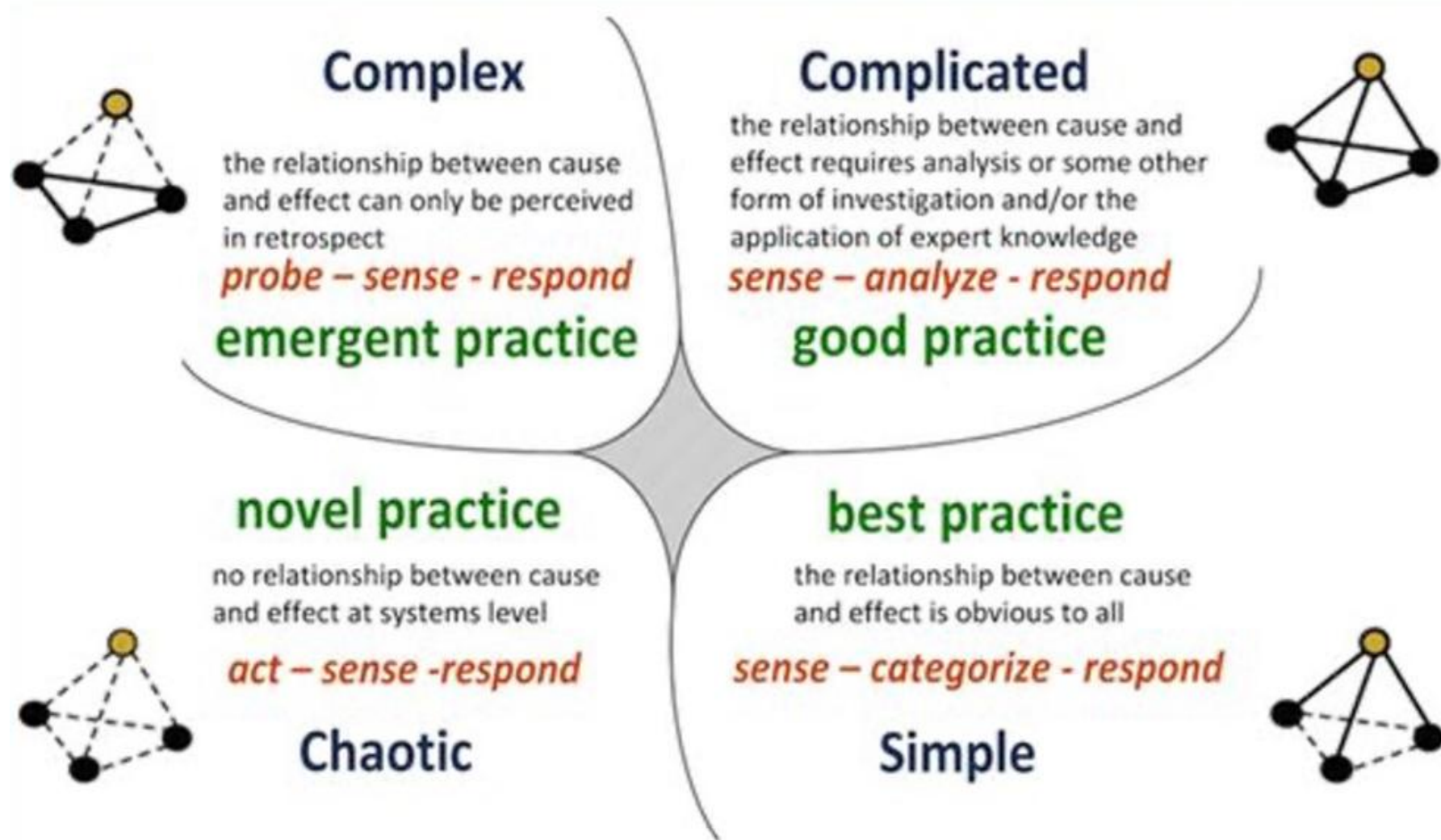
Narrow AI
(Specific Tasks)

AGI
(Human-Level Versatility)

Superintelligence
(Beyond Human)



Occupational Safety and Health as a Problem



The Cynefin Framework. Snowden, D.J. Boone, M. 2007. "A Leader's Framework for Decision Making". Harvard Business Review. November 2007. pp. 69–76.

Will there be, let's say, in 10 – 20 years from now, Occupational Health and Safety Problem?

yahoo!finance Search for news, tickers or companies  News

[My Portfolio](#) [News](#) [Tariffs Ruling](#) [Markets](#) [Research](#) [Community](#) [Personal Finance](#) [Videos](#) [Watch Now](#)

BENZINGA

Elon Musk Says Robots Will Replace 'All Jobs' And Work Will Be Optional, Like 'Growing Vegetables' Instead Of Shopping At The Store

Concluding Remarks



AI – will accelerate the creation of **new and better solutions** to the OHS Problem



LLMs performance will **keep accelerating AGI and Superintelligence** in the horizon



Whole classes of **jobs are expected to going away**



Major changes and disturbances are expected

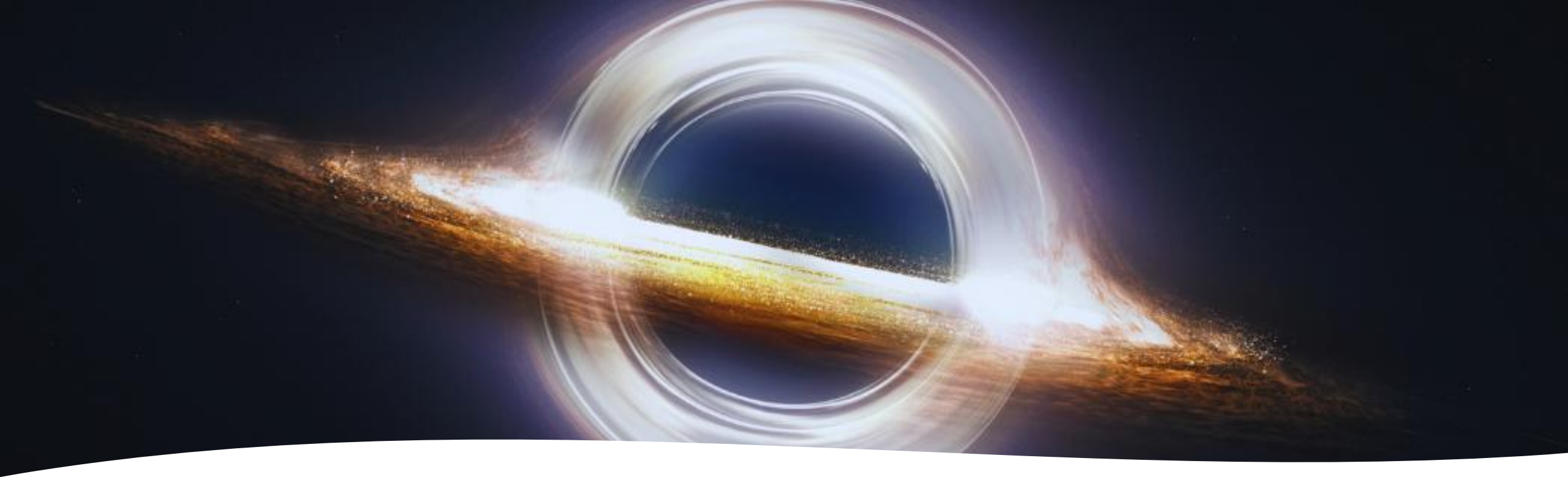
POSTHAVEN

Sam Altman

[« Back to blog](#)

The Gentle Singularity

We are past the event horizon; the takeoff has started. Humanity is close to building digital superintelligence, and at least so far it's much less weird than it seems like it should be.



Occupational Safety and Health as an Old Problem

- New types of jobs (e.g. Space exploration or any other type of exploration)
- Systems Safety, given that jobs and decisions may be made in the future predominantly by agents and robots



ΔΗΜΟΚΡΙΤΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΘΡΑΚΗΣ

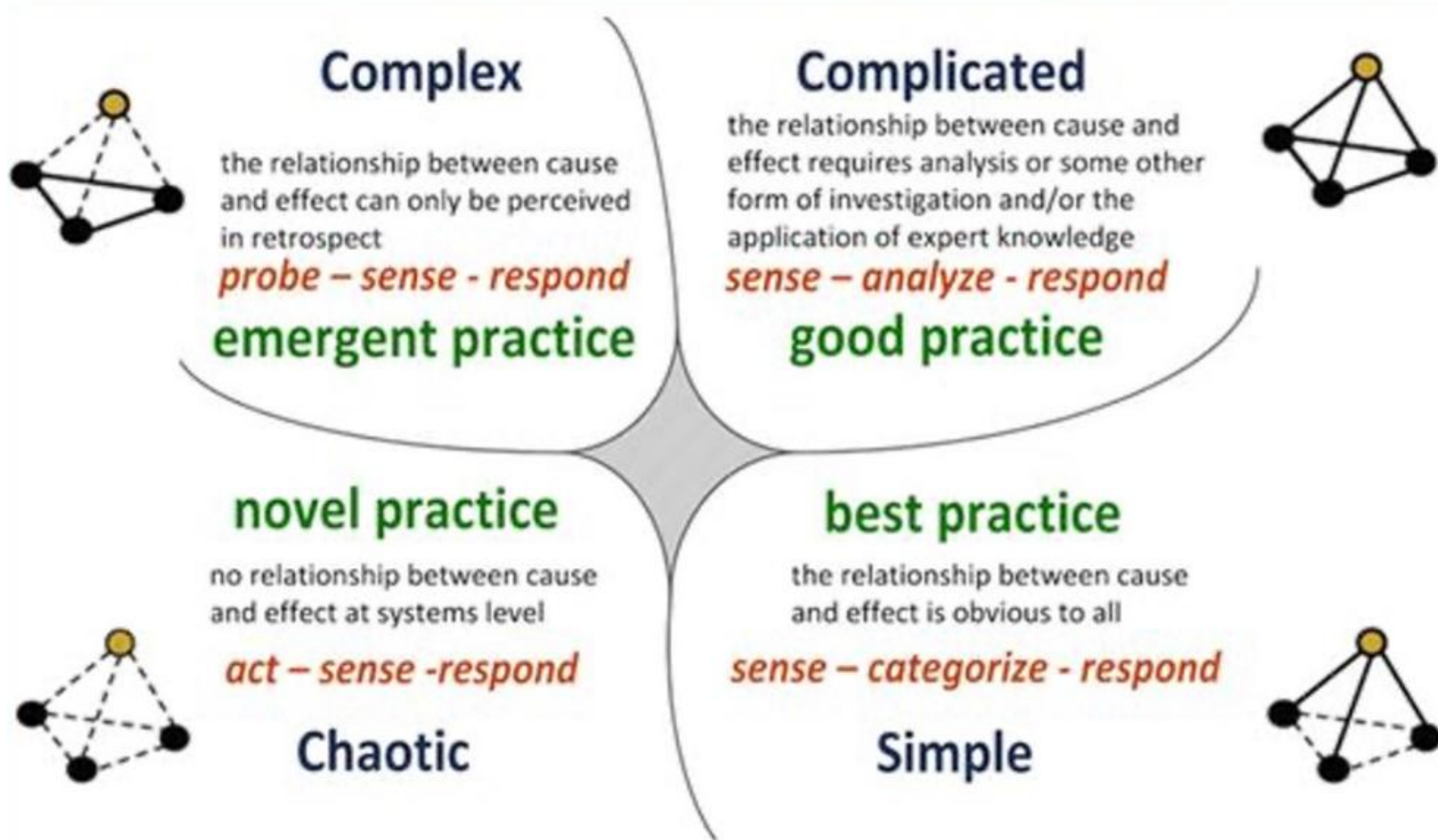
DEMOCRITUS
UNIVERSITY
OF THRACE

Ioannis M. Dokas

Professor,
Democritus University of
Thrace, Greece

Email: idokas@civil.duth.gr

Occupational Safety and Health as a Problem



The Cynefin Framework. Snowden, D.J. Boone, M. 2007. "A Leader's Framework for Decision Making". Harvard Business Review. November 2007. pp. 69–76.

