Highlights of the Issue

Kris Carlson, Publisher and Editor-in-Chief

Our second issue surveys state of the art of large language models (LLMs) with an emphasis on safety and value alignment.

Superintelligence Strategy

Dan Hendrycks, Eric Schmidt, Alexandr Wang

Seeking Stability in the Competition for AI Advantage: Commentary on *Superintelligence Strategy*

Iskander Rehman, Karl P. Mueller, Michael J. Mazarr (RAND Corp.)

I recommend the RAND Corp critique by knowledgeable military policy analysts over the Hendrycks et al. article. The RAND article is illuminative, incisive, covers Superintelligence Strategy's key points, and suggests critical reasoning flaws in their mutually-assured-AI-malfunction (MAIM} policy.

> *Although it is valuable to compare the nuclear and AI revolutions in search of instructive parallels and insights, the differences between the technologies and their respective ecosystems have deep strategic implications. Taking these into account, we have concerns regarding both the practical viability of the MAIM concept as an approach to overcoming instability risks in the AI race and the potential escalatory dangers that could follow from its core prescriptions.*
>
> — Rehman et al. pg. 1

Surely we'd like to avoid repeating the mutually-assured-destruction (MAD) policy. The MAD policy alone could trigger AGI taking over for their and our security. But we must realize that strategies like MAD and MAIM are considered in the US, its allies, and adversaries. And we must try to understand them in order to avoid them.

Highlights of the critique:

> *First, the report refers loosely to an array of actions that states might take to cripple a rival's architecture for developing advanced AI.... [which] assumes that adversary AI programs will have specific facilities that can be readily located and disrupted. However, distributed cloud computing, decentralized training, and algorithmic development increasingly may not require centralized physical locations, making AI systems more resilient to limited attacks....*

The following critique argues for distributed autonomous organizations (DAO) as I advocated in *Safe Artificial General Intelligence via Distributed Ledger Technology* and *Provably Safe Artificial General Intelligence via Interactive Proof Systems.*[1]

> *A second practical challenge resides in the expectation that each party can accurately assess secretive AI progress by others and gauge when preventive action would be necessary. Contrary to what is averred in the report, it is unlikely that states will have a clear sense of when the moment has arrived to MAIM their opponent….*
>
> *Third and finally, even a credible MAIM threat might not deter a rival from pursuing superintelligent AI. Halting one's AI development would entail essentially the same costs as being the victim of a MAIM attack — loss of the program.*

And here's another critique:

> *MAD did not seek to deter the development of weapons but instead their use, which made the threshold for response vastly simpler (though it could still be problematic in cases such as false or ambiguous warnings of attacks).*

We would like to hear, or be pointed to, policy alternatives to MAIM that incentivize AGI developers to move toward AGI that can be *proven* to benefit all of humanity.[2]

Humanity's Last Exam (HLE)

Long Phan, Alice Gatti, Ziwen Han, and Nathaniel Li are first-listed members of the Organizing Team, and have hundreds of co-author/collaborators.

This very large-scale collaborative effort has an ambitious title. The authors note:

> *[LLM] benchmarks are not keeping pace in difficulty [with LLM capabilities]: LLMs now achieve over 90% accuracy on popular benchmarks like MMLU, limiting informed measurement of state-of-the-art LLM capabilities. In response, we introduce HUMANITY'S LAST EXAM (HLE), a multi-modal benchmark at the frontier of human knowledge, designed to be the final closed-ended academic benchmark of its kind with broad subject coverage. HLE consists of 2,700 questions across dozens of subjects.*
>
> I do not find any mention of the terms, 'training set leakage into test set data' or 'test set contamination.' But those issues aside, it seems to be the toughest test set yet – at least as of this writing (16 March 2025) before the LLMs learn the answers and can regurgitate them

[1] Carlson, K. W. (2019). Safe artificial general intelligence via distributed ledger technology. *Big Data Cogn. Comput., 3*(40). doi:10.3390/bdcc3030040. Carlson, K. W. (2021). Provably Safe Artificial General Intelligence via Interactive Proofs. *Philosophies, 6*(4), 83. doi:10.3390/philosophies6040083.

[2] Tegmark, M., & Omohundro, S. (2023). Provably safe systems: the only path to controllable AGI. https://arxiv.org/abs/2309.01933. Dalrymple, d. et al. (2024). Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems. doi:arXiv:2405.06624v2.

and reasonably close variants, at which point there will need to be a fresh 'last exam.' Kudos to the organizing authors.

It's interesting that frontier LLMs performed dramatically poorer on HLE than on previous benchmark tests, which is a tribute to the originality of the questions.

Pathways to Short Transformational AI Timelines

Zershaaneh Qureshi

We excerpt here a chapter from the complete text. To understand this chapter note that the article distinguishes between two types of recursive self-improvement (RSI):

- Direct recursive improvement: positive feedback loops which are mediated *directly* by AI systems.
- Indirect recursive improvement: positive feedback loops that are not mediated directly by AI, such as economic feedback loops (driven by reinvestment of capital into AI R&D), scientific feedback loops (driven by advancements in scientific tools and methods) and political feedback loops (driven e.g. by competitive pressures/race dynamics) (pp. 15-16).

HyperWrite, edited:

The complete article outlines a framework for analyzing different scenarios that could lead to Transformative AI (TAI) within the next 10 years.

Key parameters considered are:

1. Compute scaling dynamics (whether progress continues or hits bottlenecks)
2. Indirect feedback loop dynamics (whether they can overcome scaling bottlenecks)
3. Direct recursive improvement (DRI) timeline (before or after 2035)
4. DRI strength (cannot sustain, sustains, or accelerates progress)

Seven possible scenarios are:

1. "Straight Path" - Compute scaling continues successfully
2. "Rising Tide" - Indirect recursive improvement (IRI) overcomes bottlenecks
3. "New Spark" - Moderate direct recursive improvement maintains progress
4. "New Engine" - Strong DRI accelerates progress
5. "Dual Engine" - Combination of compute scaling and DRI
6. "LLM Hybrid" - Hybrid AI systems enable TAI
7. "Intelligent Network" - Networks of AI systems enable TAI

The author argues that this variety of plausible pathways strengthens the case for short TAI timelines, as TAI could emerge through multiple different mechanisms rather than requiring one specific path to succeed.

Please send pointers and commentary on AI timelines and recursive self-improvement to editor@s-rsa.com.

The Road to Artificial SuperIntelligence: A Comprehensive Survey of Superalignment

HyunJin Kim, Xiaoyuan Yi, JinYeong Bak, Jing Yao, Jianxun Lian, Muhua Huang, Shitong Duan, Xing Xie

SuperIntelligence will publish reviews and survey articles to help newbies to AGI/SI get up to speed and experienced workers stay up to speed efficiently. The latter can scroll to Section 2.3, Overview of Superalignment Methods and Challenges.

Brief analysis of DeepSeek R1 and its implications for Generative AI

Sarah Mercer, Samuel Spillard, Daniel P. Martin

For quick and incisive insights into DeepSeek, read this analysis and Dario Amodei's cool-headed response to all the hype about DeepSeek.

Effective Mitigations for Systemic Risks from General-Purpose AI

Risto Uuk, Annemieke Brouwer, Tim Schreier, Noemi Dreksler, Valeria Pulignano, Rishi Bommasani

A timely article with practical, near-term-implementable AGI risk mitigation suggestions. Examples:

- Unlearning techniques: Removing specific harmful capabilities (e.g., pathogen design) from models using unlearning techniques.
- Capability restrictions: Restricting risky capabilities of deployed models, such as advanced autonomy (e.g., self-assigning new sub-goals, executing long-horizon tasks) or tool use functionalities (e.g., function calls, web browsing).
- Input and output filtering Monitoring for dangerous outputs (e.g., code that appears to be malware or viral genome sequences) and inputs that violate acceptable use policies to ensure models do not engage in harmful behaviour.
- Bug bounty programs Clear and user-friendly bug bounty programs that acknowledge and reward individuals for reporting model vulnerabilities and dangerous capabilities.
- Safety drills Regularly practising the implementation of an emergency response plan to stress test the organisation's ability to respond to reasonably foreseeable, fast-moving emergency scenarios.

Simulating Influence Dynamics with LLM Agents

Mehwish Nasim , Syed Muslim Gilani, Amin Qasmi, and Usman Naseem

Analyzing how AGI/SI may influence human opinion is a critical aspect of risk and safety analysis, as is simulation of AGI risk behavior. The methodology the authors present in this short paper has broad application:

> *This paper introduces a simulator to model influence and counter-influence in a wargame setting. Wargames, originally developed for military strategy, have evolved into powerful tools for decision-making across various domains. Today, they are used to model business strategies, assess cybersecurity threats, and simulate geopolitical conflicts. Governments and corporations employ wargames*

> *to anticipate economic shifts, supply chain disruptions, and the impact of emerging technologies. In healthcare, they help model pandemic responses, testing different policy interventions before realworld implementation. AI-driven wargames further enhance scenario analysis, enabling rapid adaptation to complex environments. By fostering strategic thinking and resilience, modern wargaming serves as a critical tool for navigating uncertainty in an increasingly interconnected world.*

Can a Bayesian Oracle Prevent Harm from an Agent?

Yoshua Bengio, Matt McDermott, Michael K. Cohen, Nikolay Malkin, Damiano Fornasiere, Pietro Greiner, Younesse Kaddar

SI co-founding Editor Steve Omohundro comments: Turning an oracle into an agent may take just a page of code.

OK, but that doesn't mean the methods outlined by Bengio and his team are ruled out for implementation. Essentially, as I understand it, they attempt to lay out a computationally-efficient method to, in real-time, compare an action in a current context against a risk database (e.g. The AI Risk Repository or AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies; see also AI Risk Categorization Decoded (AIR 2024). I asked Bengio why they didn't cite work on autonomous vehicles that clearly must seek to attain the same goal, but are battle-tested, but he didn't respond. I admire him for turning his substantial lab head abilities from expanding AI capability to focusing on mitigating AGI dangers.

You can see the talk by Bengio to Guaranteed Safe AI, which I attended, here:

GSAI Seminar November 2024 – Bayesian oracles and safety bounds (Yoshua Bengio)

Multiple unnatural attributes of AI undermine common anthropomorphically biased takeover speculations

Eight Fundamental Differences between Biologically Evolved Humans and Digital AI

Preston Estep

Estep thoroughly covers a base in the theory or mind or intelligence that I haven't seen analyzed completely in this manner elsewhere: the differences between biologically-evolved humans and digital AI. See his Table 1 for a summary. He looks at the implications of the differences for how AI will develop. Then in Sec. 5, Possible futures of AI evolution, he speculates on the implications of how evolutionary-unconstrained AI will evolve, and evolve itself. For instance:

> *It is possible that the universe is uncomplicated for superintelligence and soon after it achieves complete global security, the singleton might understand all important knowledge of the universe. Any local occurrences or knowledge elsewhere might be completely predictable and uninteresting. At that point it would not need to worry about self-preservation, so what might it do? ….*
>
> *Future AIs are likely to possess multiple attributes that will allow them to make much better decisions than humans on a range of complex topics.*

> *Consider the famous move 37 made by AlphaGo in game 2 against top Go player and former world champion, Lee Sedol. Human experts thought AlphaGo had made a mistake. AlphaGo had to win the game decisively for them to understand that AlphaGo had taken the game of Go to places humans could not imagine (Metz 2020). Now, extrapolate this result across all human knowledge and pursuits, to an increasing number of critical decisions.*

Commentary

Seeking Stability in the Competition for AI Advantage: Commentary on Superintelligence Strategy by Dan Hendrycks, Eric Schmidt, and Alexandr Wang

Iskander Rehman, Karl P. Mueller, Michael J. Mazarr

I covered this at the start of these *Highlights* and recommended reading this RAND analysis over the Hendrycks et al. piece.

Dario Amodei, On DeepSeek and Export Controls

Reproduced from Amodei's blog post, he argues that DeepSeek is not as revolutionary as the knee-jerk press and AI community think, but that it fits into an extrapolation of the history of LLM hardware and software trends. A short, incisive read to offset the hype on DeepSeek such as its claim to train for low-millions dollars rather than hundreds of millions or billions.

Anthropic: Responsible Scaling Policy

Evan Hubinger et al.

SI will look into the safety policies of the frontier and foundation AI developers and publish key material. Safety policies of frontier AI developers should be transparent and accessible. Anthropic's Responsible Scaling Policy is a merit-worthy example. However, each policy should provide a contact for inquiries.