Highlights of the Issue

Kris Carlson, Publisher and Editor-in-Chief

To emphasize the journal's concern with AGI safety, we inaugurate *Artificial General Intelligence* (*AGI*) by focusing the first issue on Risks, Governance, and Safety & Alignment Methods.

Risks

The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks from Artificial Intelligence

The most comprehensive AI risk taxonomy — 777 specific risks classified into 43 categories — to date has been created by workers collaborating from a half-dozen institutions. We except 11 key pages from the original 79-page report. Their 'living' Repository is online and free to download and share. The authors' intention is to provide a common frame of reference for AI risks.

Slattery et al.'s set of ~100 references is excellent and thorough. Thus, pouring through this study for your own specific interest is an efficient way to get on top of the entire current AI risk literature.

The highest of their three taxonomy levels, the Causal Taxonomy, is categorized according to the cause of the risk, Human or AI the intention , Intentional action or Unintentional, and timing — Pre-deployment of the AI system or Post-deployment. The Causal Taxonomy can be used "for understanding how, when, or why risks from AI may emerge."

They also call readers' attention to the AI Incident Database.[1] The Incident Database publishes a monthly roundup here.

AI Risk Categorization Decoded (AIR 2024)

By examining 8 government and 16 corporate AI risk policies, Zeng et al. seek to provide an AI risk taxonomy unified across public and private sector methodologies. They present 314 risk categories organized into a 4-level hierarchy. The highest level is composed of System & Operational Risks, Content Safety Risks, Societal Risks, and Legal & Rights Risks. Their first takeaway from their analysis is more categories is advantageous, allowing finer granularity in identifying risks and unifying risk categories across methodologies. Thus, indirectly they argue for the Slattery et al. taxonomy with double the categories. This emphasis on fine granularity parallels a comment made to me by Lance Fortnow, Dean of Illinois Institute of Technology College of Computing, on the diversity and specificity of human laws indicating a similar diversity may be necessary to assure AGI safety, and that recent governance proposals may be simplistic.

Indeed, Zeng et al.'s second takeaway is that government AI regulation may need significant expansion. Few regulations address foundation models, for instance. And their third takeaway is that comparing AI risk policies from diverse sources is extremely helpful to develop an overall

[1] https://incidentdatabase.ai/. "Like similar databases in aviation and computer security, the AI Incident Database aims to learn from experience so we can prevent or mitigate bad outcomes."

grasp of the issues – how different organizations conceptualize risk, for instance – and how to move toward international cooperation to manage AI risk.

AIR-Bench 2024: A Safety Benchmark Based on Risk Categories from Regulations and Policies

Applying the work just described above, Zeng et al. constructed an AI safety benchmark aligned with their unified view of private and public sector AI risk policy and specifically targeting the gap in regulation of foundation models they uncovered. They develop and test nearly 6000 risky prompts and find inconsistent responses across foundation models. Zeng et al. give examples of foundation model safety failures in response to various prompts.

This work seems a significant advance toward an AGI safety certification conducted by an AI industry consortium or an insurance company consortium along the lines of, e.g., UL Solutions (previously Underwriters' Laboratory).

A Comprehensive Survey of Advanced Persistent Threat Attribution

> *We wanted to publish this important article had to pull it due to a license conflict – please see their arXiv preprint.*
>
> *APT [Advanced Persistent Threat] attacks are attack campaigns orchestrated by highly organized and often state-sponsored threat groups that operate covertly and methodically over prolonged periods. APTs set themselves apart from conventional cyber-attacks by their stealthiness, persistence, and precision in targeting.*

This systematic review by Rani et al. of 137 papers focuses on the increasing development of automated means to detect AI and ML APTs early and identify the malevolent actors involved. They present the Automated Attribution Framework, which consists of 1) collecting the training data of past attacks, 2) preprocessing and enrichment of the training data, 3) the actual training and pattern recognition on the data, and 4) attribution — applying the trained models to identify the malevolent perpetrating actors.

The open research questions summarized by Rani et al. lead toward AI taking an increasing role in APT attribution.

Governance

Excerpts from Aschenbrenner, *Situational Awareness*

I was pointed to Leopold Aschenbrenner's 165-page missive by Scott Aaronson's blog, which said he knew Leopold during his sabbatical at OpenAI and recommended people give it a read and take it seriously. The essence of it is that ***if*** we extrapolate from recent AI progress, we will have AGI by 2030, and therefore, for national security, a Manhattan Project-style national AI effort, including nationalizing leading private AGI labs, should be mounted.

Here we reprint his Part IV, "The Project," advocating this controversial effort and describing his vision of how it will occur.

I recommend anyone concerned about the dangers of AGI, and especially those working toward AGI, read Aschenbrenner's entire book. Take a look at the Table of Contents preceding our reprint of "The Project." And we reprint his Ch. V, "Parting Thoughts," in our Commentary section.

Soft Nationalization: How the US Government Will Control AI Labs

Aschenbrenner advocates nationalizing leading AI labs into a high-security, top-secret, US federal government project. OK, how, exactly? A perfect complement to Aschenbrenner's thoughts is given by Deric Cheng and Corin Katzke of Convergence Analysis. They examine how AGI R&D nationalization could happen realistically, effectively, and efficiently. Their report outlines key issues and initial thoughts as a prelude to their own and others' detailed proposals to come. It is a beautiful piece of work, IMHO.

It is not impossible for private companies to develop AGI responsibly and securely, but the main goal of this journal is to make AGI safety the central debate in the AGI community, and the nationalized, Manhattan-style project point of view must be presented. Further, I find Aschenbrenner's arguments to be persuasive and Cheng and Katzke's thoughtful outline of how nationalization could actually occur to be convincing, e.g. (pg. 8):

> *The US may be able to achieve its national security goals with substantially less overhead than total nationalization via effective policy levers and regulation... We argue that various combinations of the policy levers listed below will likely be sufficient to meet US national security concerns, while allowing for more minimal governmental intrusion into private frontier AI development.*

Acceptable Use Policies for Foundation Models

Acceptable use policies are legally binding policies that prohibit specific uses of foundation models. Klyman surveys acceptable use policies from 30 developers encompassing 127 specific use restrictions cited in 184 articles. Like Zeng et al. in "AI Risk Categorization Decoded (AIR 2024)," Klyman highlights the inconsistent number and type of restrictions across developers and lack of transparency behind their motivation and enforcement, indicating the need to for developers to create a unified consensus acceptable use policy. The general motivations are to reduce legal and reputational risk. However, standing in the way of developers working to create a unified policy set is the motivation to use restrictions to hinder competition from using proprietary models. Enforcement can also hinder effective use of a foundation model.

Acceptable use policies can be categorized into *content restrictions* (e.g. the top 4: misinformation, harassment, privacy, discrimination) and *end use restrictions*, e.g. Anthropic's restriction on "model scraping," which is someone training their own AI model on prompts and outputs from Anthropic's model. Another use restriction is scaling up AI-created content distribution such as automated online posting.

As with the Zeng et al. articles, Klyman's article points the way to create a homogeneous acceptable use policy across a diverse AI ecosystem.

Steve Omohundro comments: "...the AI labs' 'alignment work' ... is all about the AIs rather than their impact on the world. For goodness sake, the Chinese People's Liberation Army has already fine-tuned Meta's Llama 3.1 to promote Chinese military goals! And Meta's response was 'that's contrary to our acceptable use policy!'" From the article:

> *Without information about how acceptable use policies are enforced, it is not obvious that they are actually being implemented or effective in limiting dangerous uses. Companies are moving quickly to deploy their models and may in practice invest little in establishing and maintaining the trust and safety teams required to enforce their policies to limit risky uses.*

Safety Methods

Benchmark Early and Red Team Often (Executive Summary excerpt)

Two leading methods for uncovering AI safety breaches are 1) inexpensive benchmarking against a standardized test suite, such as prompts for large language models, and 2) longer, higher-cost but more informative intensive, interactive testing by human domain experts ("red-teaming"). Barrett et al., from the UC Berkeley Center for Long-Term Cybersecurity, advocate for this two-pronged approach indicated by the article title. They analyze the methods' potential for eliminating LLM "dual" use, i.e. corrupting LLMs into creating chemical, biological, radiological, nuclear (CBRN) or cyber or other weaponry or attacks, but the methods apply to less dangerous risk testing as well.

Essentially Barrett et al. advocate frequent use of benchmarks until a model attains a high safety score, followed by intensive red-teaming to test the model in more depth and yield more accuracy. Their paraphrase of the article title is:

Benchmark Early and Often, and Red-Team Often Enough.

Against Purposeful Artificial Intelligence Failures

A paper that had to be written, and not surprisingly was, by Yampolskiy, who has sought to cover every aspect of AGI risks, is one arguing that *intentionally* triggering an AI disaster should not be entertained as an option to alert humanity to the danger of AGI.

Models That Prove Their Own Correctness

Especially in light of Dalrymple et al.'s governance proposal, Toward Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems, 'models that prove their own correctness' seems especially desirable, if not essential. Dalrymple et al. call for 1) a world model, 2) a safety specification, and 3), a means to verify the safety specification, a highly intriguing proposal, but which falls short of providing an example of such a model or means of verification (we hear that Dalrymple is working on an example).

Paradise et al. describe two uses of interactive proof systems (IPS) combined with ML to allow a model to prove its own 'correctness,' as specified by the user of the model. The first method requires access to a training set of IPS transcripts (the sequence of interactions between the Verifier and Prover) in which the Verifier accepted the Prover's probabilistic proof. The second

method, Reinforcement Learning from Verifier Feedback (RLVF; note their intentional similarity to Reinforcement Learning from *Human* Feedback, RLHF) avoids the need for the accepted transcripts (which are in essence an external truth oracle) but only after training on such a verified transcript (its 'base model') using transcript learning. From then on it can generate its own emulated verified transcripts.

The paper opens the door to other innovative applications of ML to IPS.

This is a rather deep paper that requires further analysis to judge the realization of its promise. We look forward to a revised version after its peer review at an unspecified journal. We thank Syed Rafi for the pointer to the paper and Quinn Dougherty for inviting Orr Paradise to his safe AGI reading group.

Language-Guided World Models: A Model-Based Approach to AI Control

> *Model-based agents are artificial agents equipped with probabilistic "world models" that are capable of foreseeing the future state of an environment (Deisenroth and Rasmussen, 2011; Schmidhuber, 2015). World models endow these agents with the ability to plan and learn in imagination (i.e., internal simulation)....*

Citing Dalrymple et al., Zhang et al. likewise extend the capabilities of world models to increase human control over AI. By adjusting the world model, humans can affect many context-sensitive policies simultaneously. However, for the human-AI interaction to be efficient, the world model must process natural language (NLP); hence, *language-guided* world models (LWMs). NLP also increases the efficiency of model learning by permitting them to read text. World models increase AI transparency, which NL interaction furthers by allowing humans to query models verbally.

As an example, in Sec. 5.3, "Application: Agents that discuss plans with humans," Zhang et al. describe an agent that uses its LWM to plan a task and then ask a human to review it for safety.

Commentary

Steve Omohundro, "Progress in Superhuman Theorem Proving?"

Our co-founding editor Steve Omohundro is a strong proponent of Provably Safe AI, in which automated theorem-proving will play a major role.[2] Here Steve discusses current developments in using proof to lessen LLM hallucinations, the implications of superhuman theorem-proving for safe AGI and resources for interested readers.

On Yampolskiy, "Against Purposeful Artificial Intelligence Failures"

Topic Editor Jim Miller, Professor of Economics, Game Theory, and Sociology at Smith College, critiques Roman Yampolskiy's argument against triggering a deliberate AI failure to wake the world up to AI dangers.

---

[2] Tegmark, M., & Omohundro, S. (2023). Provably safe systems: the only path to controllable AGI. arXiv. Retrieved from https://arxiv.org/abs/2309.01933.

Leopold Aschenbrenner, *Situational Awareness*, “Parting Thoughts”

Aschenbrenner dismisses his critics as unrealistic and outlines the core tenets of “AI Realism.”

Rowan McGovern, “Unhobbling Is All You Need?” Commentary on Aschenbrenner’s *Situational Awareness*

McGovern questions Aschenbrenner’s fundamental assumption that “unhobbling” alone — “fixing obvious ways in which models are hobbled by default, unlocking latent capabilities and giving them tools, leading to step-changes in usefulness” — will result in his extrapolation of recent AI progress to predict the advent of AGI by 2030. McGovern: “Unhobbling conflates computing power with intelligence.”