Highlights of the Issue: Singapore Consensus – Safety Technology

Kris Carlson, Publisher

A Note On Governance

Corin Katzke and Dan Hendrycks observe: "Since frontier AI companies are very likely to comply with the [EU] Code, securing similar legislation in the US may no longer be a priority for AI safety."[1] In other words, legal regimes do not need to repeat or reinvent regulation that is already present in major markets, but should focus on making their unique contribution to the composite global AI regulatory regime.

Articles

Singapore Consensus on Global AI Safety

Yoshua Bengio, Max Tegmark, et al.

The *Consensus* follows the First International AI Safety Report and moves the priorities identified there toward a tangible international governance framework.

> *By, adopting a defence-in-depth model, this document organises AI safety research domains, into three types: challenges with creating trustworthy AI systems (Development), challenges, with evaluating their risks (Assessment), and challenges with monitoring and intervening, after deployment (Control)., Through the Singapore Consensus, we hope to globally facilitate meaningful conversations, between AI scientists and AI policymakers for maximally beneficial outcomes., Our goal is to enable more impactful R&D efforts to rapidly develop safety and evaluation, mechanisms and foster a trusted ecosystem where AI is harnessed for the public good.*

Published under Creative Commons BY 4.0 license. Section headers were replaced with our header.

America's AI Action Plan

The *Action Plan*, just released after a few months' draft and public comment, reflects not just America's sober assessment of the near-term potential of AI to transform the world, but that of many nations – as well as of Big Tech, and of a growing number of AI startups, each of whom imagine that achieving AGI before the others could put them in a world-dominating position.

[1] Katzke & Hendrycks, Center for AI Safety, AI Safety Newsletter #59: EU Publishes General-Purpose AI Code of Practice, 15 July 2025.

In accordance with the *Singapore Consensus*, we need far-sighted individuals of the highest integrity to work toward a stable AI ecology and bring the AI arms race under control.

> *"Today, a new frontier of scientific discovery lies before us, defined by transformative technologies such as artificial intelligence... Breakthroughs in these fields have the potential to reshape the global balance of power, spark entirely new industries, and revolutionize the way we live and work. As our global competitors race to exploit these technologies, it is a national security imperative for the United States to achieve and maintain unquestioned and unchallenged global technological dominance. To secure our future, we must harness the full power of American innovation."*

- Donald J. Trump, 45th and 47th President of the United States

Outline: Proposed Zero Draft for a Standard on AI Testing, Evaluation, Verification, and Validation (TEVV)

National Institute of Standards and Technology (NIST)

This document from NIST seeks to gather information on how to adequately perform TEVV in light of the difficulties increasingly sophisticated AI systems present:

- AI systems tend to exhibit multiple levels of complexity that interfere with TEVV.
- For example, formal verification of large models is not generally achievable, and evaluation results will often reveal tendencies or likelihoods rather than definite measures.
- Testing objectives and requirements are often difficult to appropriately operationalize and measure, similar to challenges found in other areas that deal with complex concepts and systems.
- Many different entities and organizations may be performing TEVV for many different reasons.
- The quality, accuracy, coverage, and definiteness of evaluations and their results need to be contextualized and understood correctly, with appropriately managed expectations.

NIST's ultimate goal is a set of AI TEVV standards to incorporate within the existing ISO and IEC frameworks (two prominent international standards developing organizations that collaborate closely through joint committees).

In Which Areas of Technical AI Safety Could Geopolitical Rivals Cooperate?

Bucknall, Siddiqui, Bengio, Trager et al.

Supporting the key goal of international cooperation of the Singapore Consensus, Bucknall, Siddiqui et al. explore the potential for cooperation on AI safety between geopolitical rivals, focusing on risks and common-ground benefits. The goal is to guide

researchers and policymakers in fostering safe and effective international cooperation on AI safety.

Risks include advancing harmful capabilities, and especially *differentially* advancing capabilities, exposing sensitive information (e.g. strategy, infrastructure), and enabling malicious actions (e.g., tampering with systems developed collaboratively). Four areas of cooperation—verification mechanisms, protocols, infrastructure, and evaluations—are assessed for feasibility and risk. They propose that research on verification mechanisms and protocols may be particularly suitable for collaboration.

Licensed under Creative Commons BY 4.0. Headers were replaced with our header.

Standardizing Intelligence: Aligning Generative AI for Regulatory and Operational Compliance

Imperial, Jones, Madabushi

To implement the global AI safety regulatory regime toward which the Singapore Consensus moves, technical standards are needed — "established documented guidelines and rules that facilitate the interoperability, quality, and accuracy of systems and processes." Imperial et al. look at which AI safety-related standards are critical and grade the compliance capabilities of SOTA GenAI models. They describe challenges and opportunities of compliance tasks and recommend actions for developing and using standards. Computational methods are recommended to align GenAI with compliance standards.

The Evolution of AI Communication: From Chain-of-Thought to Neuralese and the Case for Interpretability Agents

Erhan Arslan

Our Editor-at-Large Gil Syswerda and others see at least a two-orders-of-magnitude speedup of LLMs communicating with each other by skipping the output layers' symbolic language and using their own internal language, called 'neuralese' by some. Arslan argues for more than that, bearing on AGI safety. He wants to evolve 'interpretability agents' to maintain human understanding of AGI behavior à la explainable AI and mechanistic interpretability. One critique from Gil, with which I agree:

> *The proposal limits AI to human-level concepts. That seems like a pretty big limitation in the race to ASI.*

In Stephen Wolfram's What If We Had Bigger Brains? post, he speculates that AGI will have a vastly greater vocabulary than ours. This means that AGI will have concepts so foreign to ours' we will not be able to understand them.