A.I. VERIFY FOUNDATION

安远 AI
CONCORDIA AI

AI Safety and Risk Management Workshop Summary

On December 3, Concordia AI and the AI Verify Foundation jointly co-hosted a closed-door workshop bringing together over 20 researchers and professionals specializing in AI technology and policy from over 6 countries. The workshop, held alongside the International AI Cooperation and Governance Forum 2024, focused on developing international consensus on AI safety testing and evaluation.

In accordance with the Chatham House Rule, participant contributions are not attributed to individuals. The readout reflects some common views but should not be taken as an endorsement by participants of the entire summary.

The participants recognized that there is an urgent need for a shared global understanding of AI capabilities, opportunities, and risks given the rapid advancement of AI capabilities. They noted that global collective action is necessary to ensure the safety of increasingly advanced systems and encourage the global community to anticipate emerging challenges and promote the safe and beneficial use of AI.

The workshop included presentations on AI safety testing approaches across various jurisdictions. Participants then broke into small groups to discuss opportunities and challenges in four key areas: risk identification, testing methodologies, risk mitigation, and ongoing monitoring of emerging risks.

Pillar 1: AI Safety Risk Identification

- To effectively address AI safety concerns, the international community should identify both current and emerging risks across multiple domains. Priority areas requiring additional research include risks from AI agents and biological design tools. Understanding which risks have already manifested versus those likely to emerge enables more strategic preparation and mitigation efforts.
- To address this and other emerging threats, some participants stated that we should establish robust early warning systems, including collaborative model testing environments and coordinated red team exercises to detect potentially dangerous capabilities.
- Given the inherent uncertainties in risk assessment, a multi-faceted approach is essential.

Pillar 2: AI Safety Risk Measurement

- Effective AI safety risk assessment requires careful matching of measurement tools to specific risk types. Current measurement approaches include static and customizable benchmarks, manual red teaming, and agent-based red teaming. Evaluations organizations would benefit substantially from sharing insights about the relative effectiveness of these tools for different risk scenarios.

- Simultaneously, it is crucial to recognize that technical benchmarks alone cannot fully anticipate how AI systems will interact with society at large. There should be complementary assessment methods that consider broader societal impacts.
- Given continued development of AI risk identification and the science of AI evaluations, formal testing standards may be premature at this stage. Instead, allowing diverse approaches to develop naturally will likely yield valuable insights that can inform future standardization efforts.
- To advance the field, the global academic community would benefit from a formalized research agenda, potentially supported by funded prizes for solving specific AI safety testing challenges. This structured approach would help focus research efforts and accelerate progress in key areas.
- A particularly promising opportunity for international collaboration is jointly developing open assessment suites. These could serve as practical tools for implementing regulatory frameworks such as the EU AI Act, while fostering greater international interoperability in AI safety testing.

Pillar 3: AI Safety Risk Mitigation

- Risk assessment bodies and AI companies should develop clear thresholds that trigger enhanced safety mitigation measures. These thresholds can be established through literature reviews and consultation with subject matter experts and relevant government authorities. When evaluations indicate breaching of relevant thresholds, developers must implement interventions based on the severity level identified.
- A defense-in-depth strategy across the AI development and deployment lifecycle is essential for comprehensive risk mitigation. While mitigation during the pre-training phase is challenging due to potential impacts on model capabilities, the post-training phase offers several viable options, including unlearning and refusal fine-tuning. At the application layer, additional safeguards such as input/output filtering and enhanced refusal mechanisms can provide further protection.
- The dual-use nature of many AI domains presents particular challenges, for example, restricting models' access to general biology knowledge could potentially impede beneficial innovation. While a consensus on this challenge remains elusive, one potential approach would be to create whitelists of vetted actors who can access models with specific knowledge capabilities under a know-your-customer (KYC) scheme.
- Ensuring security of models against theft and interference is also an important and neglected mitigation. Mitigation of potential risks from open-weight models remains an open problem, and one potential approach is preventing the model from being fine-tuned for malicious purposes.

Pillar 4: AI Safety Risk Monitoring

- Global perceptions of AI risks are fundamentally compatible, though factors including institutional differences influence which specific risks receive emphasis in different regions. The Bletchley Declaration demonstrates broad consensus on the importance of addressing public safety risks from general-purpose AI systems, even as complete consensus on a single risk taxonomy remains challenging.
- AI safety governance operates across multiple levels, from corporate to national to international layers. A tiered approach enables appropriate distribution of responsibilities: national governments can implement domestic risk assessment and reporting requirements, while international cooperation can focus on establishing safeguards against catastrophic risks and defining clear red lines.
- The challenge of incident reporting requires careful consideration of incentive structures. Overly punitive responses to safety incident reports may discourage corporate transparency and create unintended chilling effects. To promote open reporting, regulators should balance accountability measures with positive incentives such as third-party ratings or certifications, preferential capital access, and inclusion in a consortium for trusted providers. Consequences for non-reporting might include corporate liability, executive accountability, and reputational impacts.