International Network
of AI Safety Institutes
NOVEMBER 20–21, 2024 | SAN FRANCISCO

Joint Statement on Risk Assessment of Advanced AI Systems

International Network of AI Safety Institutes
November 20, 2024

Assessing the risks of advanced AI systems, which include systems referred to as frontier AI systems, dual-use foundation models, general-purpose AI models, and advanced generative AI systems, presents novel challenges. Advanced AI systems have capabilities across a broad range of contexts, enabling them to be used and misused, accidentally or intentionally, in ways that can be difficult to predict, measure, and mitigate. Addressing these challenges is core to the mission of the International Network of AI Safety Institutes.

Following commitments made in the Bletchley Declaration and the Seoul Statement of Intent, as well as the progress made through the OECD, G7 Hiroshima Process, the Frontier AI Safety Commitments, and other relevant initiatives, in this document the International Network of AI Safety Institutes highlights **six key aspects of risk assessment of advanced AI systems**.

The Network is committed to building on these six key aspects to establish a shared scientific basis for risk assessments of advanced AI systems. This may involve conducting joint risk assessments and cooperative scientific research, recognizing that the science and practice of advanced AI risk assessment continues to evolve. Individual network members retain flexibility to conduct, apply, and adapt any risk assessments or risk-benefit trade-offs in line with international and domestic frameworks.

Key aspects of risk assessment of advanced AI systems:

1. Actionable

Risk assessments should be carried out in a manner that can directly inform proportionate and effective mitigation measures, for example by estimating risk in relation to specified evaluation criteria, such as tolerance levels or thresholds. Risk assessments may target prioritized risk domains, which should be defined with input from a range of stakeholders, including subject matter experts. Risk domains can be prioritized according to multiple criteria, including their severity, likelihood of occurrence, or the level of societal resilience in that domain. Risk tolerance can be defined in advance to help translate the level of identified risk to a particular set of mitigation measures. Example scenarios in which risk tolerance would be met can be used to illustrate this link between evaluation criteria and risk tolerance.

2. Transparent

Risk assessments should to the greatest extent possible be transparent in their methodology and results. Transparency can help ensure that risk assessments are evidence-based, interpretable, and consistent. Transparent risk assessments provide benefits to industry, academia, civil society, and the public, as they foster greater understanding of how and why risk assessments are conducted, as well as the potential risks found. Specific requirements around transparency may differ according to the nature of the assessed risk, the organization responsible for the risk assessment, and the sensitivity of the information related to the assessment. Disclosure and handling of sensitive information should account for safety, privacy, and commercial considerations.

3. Comprehensive

Risk assessments should be comprehensive and connected to a broad range of potential and existing real-world harms through the use of a variety of assessment methods. An estimation that an advanced AI system may pose a particular risk should map to the potential impact of the risk if it manifests as part of a system deployed to users. The estimation should also consider the downstream impacts from adoption of advanced AI systems at scale. These estimations should be evidence-based, falsifiable and balanced. They should take into consideration how real-world harms can be avoided and how advanced AI systems might address risks posed to solving some of the world's greatest challenges, such as the achievement of the Sustainable Development Goals.

4. Multistakeholder

Risk assessments should be multistakeholder in their approach and in the interpretation of results. The scale and increasing impact of advanced AI systems demands a more integrated ecosystem of AI safety that includes diverse disciplines, perspectives and experiences, including from across the AI lifecycle. The views of stakeholders from the private sector, civil society, academia, government, and other sectors should be considered when carrying out risk assessments. Experts from computer science, engineering, and mathematics should collaborate with experts from disciplines relevant to the risk assessed, including the social sciences and other fields, to better assess the risks of advanced AI systems. This can also be important for identifying any trade-offs among risks. This collaborative process should enhance robustness of risk assessments by involving the most relevant stakeholders, including those who may be directly impacted.

5. Iterative

Risk assessments should inform concrete decisions and be conducted at regular intervals to adapt to progress in advanced AI systems and AI safety research. Risks should be assessed and mitigated across the AI lifecycle, as appropriate, including before advanced AI systems are deployed and throughout the development process in an iterative, holistic manner. Risks and harms should also be monitored post-deployment to account for flaws and vulnerabilities that emerge as advanced AI systems are integrated into products and services, including the risk associated with downstream misuse and interaction with other deployed systems. Information from ongoing monitoring should be incorporated into risk assessment processes.

6. Reproducible

Risk assessments should be, to the extent possible, reproducible and appropriately documented. Methodologies and results that are reproducible allow for independent third-party evaluators to replicate, verify and validate risk assessments. This can help improve risk assessment processes, decrease error in results, and increase interoperability among methods and actors.