AI Safety Forecast Report

Yeongkyun Jang, Minnseok Choi, Jiyeon Cho, Kyungho Song
AI Safety Policy and Strategic Cooperation, Korea AI Safety Institute

1. Overview

This *AI Safety Forecast Report* presents a future-oriented analysis of global discourse on AI safety, grounded in 98 key news articles selected between May and July 2025. Curated by PhD-level experts in AI safety, these articles highlight the most critical developments and debates shaping the field. From this dataset, 249 unique or overlapping keywords (nodes) and 970 associated links were extracted to conduct a comprehensive network centrality analysis.

Using network centrality metrics such as weighted-degree centrality, eigenvector centrality, and betweenness centrality, a group of doctoral-level policy researchers at Korea AI Safety Institute (K-AISI) developed three plausible future scenarios. These scenarios include the institutionalization of global norms, the expansion of industry-led safety testing frameworks, and the securitization of AI governance. The report presents a strategic forecasting framework that guides proactive and collaborative approaches to AI safety in an increasingly complex and competitive global environment.

2. AI Safety Trends Driven by Network Centrality Analysis [See Appendix A, B]

2.1 Rise of a Risk-Oriented AI Safety Framework

The concept of *risk* stands at the center of conversations surrounding AI safety. Among all related terms, risk ranks the highest across three key network centrality measures—weighted-degree, eigenvector, and betweenness centrality—highlighting its dominant role in the discourse. This indicates that risk serves as a foundational pillar in how AI safety is framed, encompassing not only technical malfunctions or performance shortcomings but also broader societal, economic, and security threats.

This emphasis on risk has intensified in response to recent advancements, including the widespread deployment of generative AI, the rise of massive foundation models, and growing interest in artificial general intelligence (AGI). As AI systems become more capable of perceiving, reasoning, and making decisions autonomously, the potential dangers associated with them become increasingly complex and difficult to foresee. Consequently, AI safety is no longer seen as optional—it is now regarded as an essential principle that must evolve in step with ongoing technological progress.

2.2 The Rise of AI Safety Governance and the UK's Emerging Leadership

As AI safety becomes a central focus in policy debates and regulatory frameworks, key governance actors are taking on increasingly prominent roles within the global landscape. Centrality analysis highlights *AISI* and *UK* as highly influential terms alongside *risk*, indicating that global AI safety institutes are playing a central role in shaping the discourse. The United Kingdom, in particular, has emerged as a leader—not only through technical initiatives but also by advancing institutional coordination and strengthening diplomatic leadership to promote international collaboration on AI safety.

2.3 Norm-Setting and Institutionalization: Law, Standards, and Evaluation

As AI safety becomes more firmly embedded in policymaking and legislation, the terms *act (law)*, *evaluation*, and *standard* have emerged as central tools for implementation. The high rankings of *act* in both eigenvector and betweenness centrality reflect the rapid rollout of legal frameworks like the EU AI Act.

On the foundation of such legal structures, *evaluation* and *standard* function as practical mechanisms to ensure compliance. With increasing emphasis on risk-based approaches, requirements such as impact assessments and model validation are becoming legally mandated, while efforts to establish technical standards are advancing in parallel. The strong centrality of standard underscores the critical role that standardization plays in bridging the gap between fast-moving technologies and effective regulatory policy in AI safety.

2.4 Advancing Technology and Emerging Uncertainties: LLMs and Frontier AI Firms

On the technical front, *Large Language Models (LLMs)* stand out as the most prominent example of the complexities surrounding AI safety. The term LLM ranks among the top across all network centrality metrics, underscoring its central role in current debates. These models have demonstrated impressive expressive power through scaling, yet they also introduce significant risks—ranging from loss of control and misinformation generation to the potential triggering of dangerous behaviors.

Keywords like *Anthropic* and *test* point to two key developments: the rise of company-led safety validation efforts and the evolution of tools for early risk identification. At the same time, the keyword test reflects the broader adoption of risk analysis methods like red teaming, adversarial testing, and stress testing, which help uncover vulnerabilities before deployment. These technical strategies highlight the importance of pairing engineering-level safety mechanisms with policy-level oversight. Ultimately, such tools are gaining recognition as essential, data-driven resources that support proactive decision-making and contribute meaningfully to the overall assurance of AI safety.

2.5 Normative Competition: The Tri-Polar Landscape of the EU, U.S., and China

AI safety has become a central issue in global governance, with *EU*, *US*, and *China* emerging as key actors competing to shape international norms and technical standards. The EU is leading with a regulation-first approach through the AI

Act, which formalizes a risk-based framework and aims to export a European model rooted in legal obligations and human rights principles.

In contrast, the United States is pursuing a more flexible, industry-driven model focused on voluntary standards and innovation, supported by NIST (CAISI) and the 2025 AI Action Plan. Meanwhile, China is advancing a state-centered governance system based on control and social stability, having introduced early regulations on generative AI. These three distinct approaches reflect an ongoing global competition over the direction of AI governance.

3. Key Points from Expert Group Discussions in the Scenario Development Process

Through network centrality analysis, the key keywords (*risk*, *act*, *standard*, *UK*, *LLM*, *Anthropic*, *test*, *EU*, *US*, *China*, etc.) were interpreted by experts from three perspectives. Keywords with high weighted degree were seen as immediate impact factors with strong current frequency and connectivity, while those with high eigenvector centrality were identified as hubs shaping medium-term trajectories through their links to other influential keywords. Keywords with high betweenness centrality were regarded as strategic connectors bridging different clusters in the network. This multi-dimensional interpretation provided a structural understanding of the roles and persistence of each keyword within the AI safety discourse.

Based on this analysis, experts grouped the keywords into three thematic clusters: policy and governance (*risk*, *act*, *standard*, *UK*, *EU*), industry and technology (*LLM*, *Anthropic*, *test*, *research*), and geopolitics and security *(US*, *China*, *security*). Each cluster served as a narrative axis for potential scenario building. Applying the futures studies principle of "structural stability," experts judged that highly central keywords were likely to exert sustained influence in the near future despite short-term fluctuations. Consequently, these terms were designated as key driving factors for the scenario forecasting process.

Using these key driving factors, experts designed a scenario framework. Major keywords from each cluster were established as central axes, and their potential to trigger policy changes, industry responses, and geopolitical ripple effects was mapped. Interactions and possible combinations among keywords were evaluated to identify realistic pathways. This process produced three alternative futures: the "Transition to an Era of Norm-Setting" dominated by policy and governance (Scenario A), the "Expansion of Industry-Led Safety Testing Frameworks" driven by the technology sector (Scenario B), and the "Geopoliticization and Securitization of AI Safety" characterized by intensified global rivalry (Scenario C). Each scenario's likelihood and associated risks were then assessed in detail.

4. Scenario Forecasts Based on Key Driving Factors

4.1 Scenario A. Transition to an Era of Norm-Setting

This scenario is centered on key terms such as *risk*, *act*, *evaluation*, *standard*, *UK*, and *EU*, depicting a global landscape where AI safety governance enters a full-scale phase of institutionalization and normative regulation. As AI systems

grow in scale and their societal and economic impacts deepen, the concept of risk becomes the central organizing principle of global governance. In response, major countries utilize legal frameworks, pre-deployment evaluation systems, and technical standards as strategic tools to strengthen their roles as rule exporters in the international AI market.

In particular, The European Union takes the lead by implementing advanced evaluation frameworks, notably the EU AI Act, which serve as reference models for countries building their own regulatory systems. In the private sector, it becomes common practice for companies to conduct publicly guided safety assessments before deploying LLMs or autonomous systems. This scenario promotes stronger international coordination and allows for proactive risk management, but it may also intensify regulatory competition and risk slowing innovation due to overregulation.

4.2 Scenario B. Expansion of Industry-Led Safety Testing Frameworks

This scenario, built around keywords such as *Anthropic*, *test*, *LLM*, and *research,* envisions a future in which Frontier AI companies, rather than governments or regulators, take the lead in addressing AI safety. Companies like Anthropic, OpenAI, and Google DeepMind develop their own safety evaluation frameworks and testing protocols, which are shared with the broader global community. These company-led initiatives gradually evolve into de facto international standards, especially as they prove more agile and immediately applicable than traditional regulatory processes. Techniques such as red teaming, adversarial testing, capability evaluation, and alignment testing become widely adopted and standardized, laying the foundation for a robust private-sector safety infrastructure that increasingly interacts with public institutions.

The strength of this scenario lies in its flexibility and speed, enabling rapid response to emerging risks while supporting continued innovation in AI development. However, it also brings significant challenges: the dominance of private standards could lead to imbalanced norm-setting, while public oversight may be weakened. These risks highlight the need for strong accountability mechanisms and external checks to ensure that safety frameworks remain transparent, inclusive, and aligned with the public interest.

4.3 Scenario C. The Geopoliticization and Securitization of AI Safety

This scenario builds on keywords like *US*, *China*, and *security*, along with other concepts showing high betweenness centrality, to describe a future where AI safety becomes a key domain of geopolitical rivalry and national security planning. The United States and China each develop their own AI safety certification frameworks and enter a non-recognition regime, where they no longer accept the validity of one another's standards. As global mechanisms for sharing risk information and incident reports begin to erode, states adopt increasingly inward-looking and protectionist approaches, citing the need to defend national interests and ensure control over strategic technologies.

In such a world, safety standards are shaped by each country's foreign policy objectives, surveillance systems, and national security concerns. This divergence leads to a fragmentation of global AI governance and weakens the trust-based interoperability required for international cooperation. Even among allied nations, political friction may intensify over decisions related to AI system certification and the cross-border flow of AI-enabled services. While this scenario

may strengthen technological sovereignty and domestic control, it also risks undermining global collaboration, widening information asymmetries, and reducing the efficiency of collective responses to AI-related risks.

5. Strategic Directions

As AI systems continue to scale in both size and complexity, the global community faces an urgent need to establish a collective framework for addressing AI-related risks. To prepare for a more regulated future, it is important to reach international agreement on shared definitions of risk, aligned evaluation criteria, and consistent safety testing procedures. Ensuring compatibility among existing frameworks like the EU AI Act. At the same time, the framework must be inclusive, allowing developing countries to participate meaningfully in shaping global AI governance.

Given the potential for geopolitical division and intensifying technological competition, countries should take early action to build systems for sharing risk-related information based on mutual trust. This includes creating an international platform that facilitates the anonymized and publicly motivated exchange of data such as incident reports, testing outcomes, red team results, and known vulnerabilities. Such a mechanism would help support early warning capabilities and enable coordinated global responses. In addition, international collaboration between public and private sectors should focus on building decentralized testing infrastructure and providing low-cost, easily accessible safety assessment toolkits. These efforts are vital to ensure that all AI developers meet minimum safety requirements and to prevent the global fragmentation of AI safety standards. Ultimately, they contribute to the development of a trustworthy and cooperative international AI environment.

Appendix A. Results of Network Centrality Analysis

Rank	Weighted-degree centrality		Eigenvector centrality		Betweeness centrality	
1	risk	68	risk	1	risk	6520.111
2	uk	68	aisi	0.913393	uk	4530.462
3	aisi	68	uk	0.760775	act	4354.25
4	act	56	act	0.710344	aisi	4186.211
5	llm	44	research	0.69209	llm	3563.522
6	research	41	test	0.568686	research	3053.155
7	eu	40	eu	0.534783	anthropic	2839.165
8	us	36	evaluation	0.532874	us	2821.461
9	evaluation	32	standard	0.527588	china	1712.166
10	anthropic	32	llm	0.511838	security	1689.368
11	china	30	security	0.504689	report	1536.732
12	test	28	us	0.486561	test	1524.555
13	standard	28	china	0.464688	eu	1441.522
14	security	28	anthropic	0.409727	responsibility	1375.04
15	report	24	framework	0.389089	gpt	1364.327
16	transparency	18	assessment	0.38897	canada	1053.843
17	system	16	openai	0.349065	transparency	1038.24
18	singapore	16	system	0.335531	openai	951.0013
19	assessment	16	frontier ai	0.334059	evaluation	945.0715
20	misuse	16	transparency	0.331106	assessment	852.4142

Appendix B. Network Topology