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* examines how global discussions on AI safety have evolved by analyzing 125 major news articles published between August and October 2025. From these articles, researchers at Korea AI Safety Institute (AISI) identified 232 core keywords and mapped 849 links between them, creating a detailed picture of how ideas and actors cluster within the current debate. The analysis shows that terms such as *risk*, *regulation*, *standard*, *framework*, *the United States*, *China*, *LLM*, and *Anthropic* occupy central positions across the network. These findings suggest that conversations about AI safety are no longer fragmented. Instead, they are becoming more structured, with technical, policy, and geopolitical themes increasingly intertwined. This provides a strong empirical foundation for understanding the forces that are reshaping global AI governance.

Using the patterns revealed through network centrality analysis, the research team developed three scenarios that outline where AI safety may be headed next. The first scenario envisions a world where regulation, standards, and formal oversight become the backbone of global AI governance. The second scenario describes a future in which frontier AI companies take the lead by advancing their own testing methods and safety frameworks. The third scenario captures a landscape where geopolitical tensions deepen and countries move in increasingly different directions on AI safety. Although each scenario highlights a distinct pathway, they are closely connected through shared trends, such as the rise of cohesive safety frameworks and the growing prominence of agentic AI systems. Taken together, these scenarios underscore the importance of coordinated international action to navigate a rapidly changing technological and political environment.

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

2.1 U.S.-Centered Reconfiguration of the Global AI Safety Landscape

The centrality analysis shows a clear realignment in the global AI safety landscape. The United States now occupies the most dominant position, ranking first in both weighted degree and betweenness centrality. This means the U.S. is

* Please refer to the "AI Safety Forecasting Methodology," which is posted on the website of Korea AI Safety Institute (K-AISI), for the research methodology that logically underpins this report: https://www.aisi.re.kr/kor/article/ATCL75b4fb0a5/67?mno=&pageIndex=1&searchCondition=&searchKeyword=

functioning as the primary hub connecting technical research, regulatory initiatives, and industrial developments. Earlier network structures placed more emphasis on concepts such as *risk*, *AISI*, and the *UK*, but the current data reflects the influence of the 2025 AI Action Plan, CAISI's expanded role, and the leadership of American frontier AI companies.

The exceptionally high betweenness score indicates that the U.S. is deeply integrated into discussions across both technological and policy domains. Rather than relying solely on formal regulation, the United States appears to guide global AI safety efforts through a combination of innovation driven incentives and flexible governance tools. This dual approach positions American institutions and companies as key agenda setters in shaping the next phase of AI safety governance.

2.2 The Dominance of Frameworks and the Systematization of AI Safety

Another major change in the updated network is the rise of *framework* as the most influential node in eigenvector centrality. Its prominence indicates a shift away from narrow, tool-specific debates and toward broader questions about system architecture, governance models, and procedural safeguards. Frameworks now serve as the organizing principle for safety activities that were previously scattered across guidelines, audits, assurance methods, and risk-management tools.

This trend reflects the growing need for more coherent and structured governance approaches as AI models become more capable and more widely deployed. With risks evolving quickly and across multiple dimensions, frameworks help unify legal requirements, technical evaluations, and operational practices into a single safety ecosystem. Their rise suggests that AI safety is entering a more mature phase, one defined not just by individual technical interventions but by fully articulated governance structures that can adapt to new challenges.

2.3 Consolidation of the Regulation and Safety Axis

The latest results also highlight the strengthening connection between *regulation* and *safety*. Regulation performs strongly across all metrics and ranks second in betweenness centrality, while safety appears consistently as one of the most influential nodes in the network. Together, their prominence suggests that global AI safety governance is increasingly organized around legally grounded responsibility structures and forms of systematic oversight.

This shift is reflected in how governments are embedding safety expectations into legislation and requiring industry participants to implement formal evaluation and transparency practices. Technical assessments, standards development, and risk mitigation measures are increasingly intertwined with regulatory decisions. As the regulation and safety axis becomes more established, it is shaping a more integrated governance landscape that includes not only the European Union but also the United States, China, and other emerging regulatory actors.

2.4 Expanding Influence of Frontier AI Firms in Safety Governance

Frontier AI firms appear more prominently in the updated network. Companies such as *Anthropic*, *Meta*, *OpenAI*, and the developers of *Claude* have high scores across multiple centrality measures. Their influence reflects the fact that much of the practical work in AI safety, including red teaming, safety tuning, evaluation design, and operational risk assessment, is being carried out within private laboratories rather than government agencies.

This development is creating a hybrid governance environment in which private firms play an increasingly influential role in shaping expectations and norms. Their internal frameworks, testing methods, and safety procedures often move faster than public standards, and as a result they help set benchmarks that international actors reference and adopt. This growing influence suggests that authority in AI safety is shifting toward industry driven infrastructures, making these companies central to the formation of global safety norms.

2.5 Emergence of Agentic AI and a Broader Technical Risk Profile

The updated network also highlights the rising importance of technical concepts such as *GPT*, *agent AI*, *chatbot*, and *ai system*. These terms appear near the top of the analysis, with chatbot standing out due to its particularly high betweenness score. This pattern reflects growing attention to the risks associated with more interactive and autonomous systems, especially those capable of taking actions, making decisions, or operating independently across a variety of contexts.

The increasing centrality of *agent AI* suggests that concern is shifting from static language models to more dynamic systems capable of planning, multi-step reasoning, and delegated task execution. These concepts are closely linked to evaluation, security, and legislation, indicating that policymakers and researchers view agentic AI as a major new area for safety governance. Taken together, these developments point toward the early stages of a major paradigm shift in AI, one in which ensuring the safety of autonomous systems becomes a central challenge for institutions and developers.

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

The expert group began by emphasizing that the key driving factors identified through network centrality analysis are the backbone of the entire scenario development process. Keywords such as *risk*, *regulation*, *standard*, *framework*, *the United States*, *China*, *the United Kingdom*, *LLM*, *Anthropic*, *test*, and *safety* consistently ranked high across multiple centrality measures. Because these terms cut across policy discussions, technical debates, and geopolitical narratives, the experts viewed them as stable forces that are likely to shape AI safety well beyond short-term news cycles. This understanding reinforced the idea that scenarios should be grounded in data driven indicators that show persistent influence rather than speculative or episodic trends.

Based on these insights, the experts organized the key driving factors into three main thematic areas: policy and governance, industry and technology, and geopolitics and security. Each group of factors pointed to distinct pathways

through which global AI safety might evolve, yet all were closely interconnected. For example, terms like *risk*, *regulation*, *standard*, and *framework* reflected a growing move toward more formal oversight structures. The presence of *LLM*, *Anthropic*, *test*, *GPT*, and *agent AI* highlighted the expanding influence of frontier AI firms and the challenges brought by increasingly autonomous systems. Meanwhile, the prominence of *the United States*, *China*, *security*, and *AI system* pointed to the geopolitical pressures shaping international cooperation and fragmentation. These three thematic areas became the narrative foundation for building the scenarios, helping experts explore how developments in one area could reinforce or limit developments in others.

Using these organized clusters and their most influential driving factors, the experts developed three plausible scenario directions for global AI safety. The first scenario focuses on governance led consolidation, where regulations, standards, and institutional frameworks take center stage. The second scenario highlights industry led systematization, where testing methods and evaluation tools created by frontier AI companies play an increasingly important role in shaping global expectations. The third scenario captures a world where geopolitical competition grows stronger and AI safety becomes entangled with national security concerns. Across all three scenarios, the experts noted that the interactions among the key driving factors, especially the rising importance of frameworks and regulation and the emergence of agentic AI, will ultimately determine whether one scenario dominates, whether they overlap, or whether new hybrid pathways emerge.

4. Scenario Forecasts Based on Key Driving Factors

4.1 Scenario A. Governance Led Consolidation Around Regulation and Safety

This scenario is grounded in key driving factors such as *regulation*, *safety*, *standard*, *framework*, *the European Union*, and *China*. Together, they describe a future in which AI safety governance becomes more structured and institutionally coordinated. As AI capabilities expand and agentic AI systems raise new technical and security concerns, governments place increasing emphasis on regulation and safety as the core concepts guiding their approach. Countries strengthen their legal and oversight systems through evaluation requirements, transparency rules, and standardized risk management procedures.

In this environment, the European Union maintains its position as a global reference point through the implementation of the EU AI Act and its associated assessment mechanisms. China continues developing its own governance architecture with updated versions of its AI Safety Governance Framework, offering a state-driven model centered on national security and strong administrative oversight. Other countries begin aligning themselves with one of these regulatory ecosystems, adopting shared practices such as pre-deployment testing, incident reporting, and documentation standards. Although this scenario enhances consistency and strengthens global risk governance, it may also increase regulatory burdens and create competitive pressures that could slow rapid innovation.

4.2 Scenario B. Expansion of Industry Led Systematization and Safety Frameworks

This scenario is shaped by driving factors such as *framework*, *test*, *evaluation*, *Anthropic*, *GPT*, *Meta*, and *OpenAI*. It envisions a future where frontier AI companies play a leading role in building and refining the practical systems that underpin AI safety. These companies develop detailed internal processes for red teaming, adversarial testing, capability evaluations, and ongoing monitoring. Over time, these processes evolve into structured assurance frameworks that begin to influence safety expectations beyond the private sector.

Companies such as Anthropic, Google DeepMind, Meta, and OpenAI share their methods with researchers, auditors, and public institutions, creating tools that are modular enough to be adopted in different regulatory contexts. This results in a hybrid ecosystem where public authorities set broad principles while private frameworks provide the technical detail needed for implementation. The strength of this scenario comes from its ability to adapt quickly to new risks, especially those related to agentic AI and highly interactive chatbot systems. However, it also presents challenges: the concentration of influence within a small number of companies can limit transparency and fairness, and smaller developers may struggle to keep pace. Ensuring accountability and maintaining openness becomes essential to prevent the private sector from shaping safety norms without sufficient public oversight.

4.3 Scenario C. Geopolitical Fragmentation and Security First AI Safety

This scenario is driven by key factors including *the United States*, *China*, *security*, *cyber security*, *weapon*, and *legislation*. It describes a future where AI safety becomes tightly linked to geopolitical rivalry and national security strategies. Both the United States and China develop their own safety certification systems and begin incorporating them into export controls, alliance technology rules, and strategic industrial policy. As competition intensifies, the two countries gradually move toward a situation in which they no longer accept one another's safety standards.

As this dynamic unfolds, global mechanisms for sharing incidents, vulnerabilities, and other risk related information begin to weaken. Nations adopt more protectionist and security focused approaches, motivated by concerns about critical infrastructure threats, and the misuse of autonomous systems. Even among allied countries, disagreements grow over which certification pathway to follow and how to manage the international flow of AI systems and services. This fragmentation undermines trust and makes cooperation more difficult, slowing the creation of unified global standards at a time when advanced and agentic AI systems require more coordination rather than less. While the scenario strengthens domestic control of sensitive technologies, it can also widen information gaps and weaken collective preparedness against emerging AI risks.

5. Strategic Directions

The analysis presented in this report shows that global AI safety is being shaped by three powerful and overlapping forces: the consolidation of regulation and safety as core governance principles, the growing influence of industry led safety frameworks, and the increasing weight of geopolitical competition. As these dynamics unfold, the international

community faces a pressing need to establish a coordinated foundation for managing AI related risks. In a landscape where regulatory systems are diversifying, frontier companies are moving faster than public institutions, and major powers are competing to shape global norms, it becomes essential to create structures that preserve interoperability and trust while allowing room for national and technological differences.

To move in this direction, countries should work toward harmonizing definitions of risk, aligning evaluation criteria, and creating more consistent approaches to safety testing. A shared baseline for concepts such as transparency expectations, pre deployment assessments, and incident reporting would make it easier for different governance models to coexist without fragmenting global markets. At the same time, the world needs to invest in practical safety infrastructures that support both large and small developers. Decentralized testing facilities, open and low cost assessment toolkits, and credible third party assurance ecosystems can help ensure that safety expectations remain accessible and do not become concentrated in the hands of only a few major companies. These efforts will be especially important as agentic AI systems introduce new and more complex technical risks.

Finally, the rise of geopolitical tension makes it crucial to establish trusted channels for sharing risk related information. Even under conditions of partial trust, countries need mechanisms that allow them to exchange anonymized incident reports, vulnerability data, red team insights, and evaluation findings without compromising national security. Such systems would help maintain early warning capabilities and support coordinated responses at a time when global cooperation is becoming harder to sustain. Regional partnerships, multi stakeholder collaborations, and shared public private initiatives can further reinforce this resilience. By investing early in these forms of cooperation, the international community can strengthen its preparedness for emerging AI risks and maintain the foundations of a trustworthy and inclusive global AI environment.

Appendix A. Results of Network Centrality Analysis

Rank	Weighted-degree centrality		Eigenvector centrality		Betweeness centrality	
1	us	60	framework	1	us	9505.703
2	risk	44	test	0.999017	regulation	5454.051
3	anthropic	40	risk	0.868198	framework	4189.16
4	regulation	32	us	0.713746	risk	3127.906
5	safety	32	safety	0.640097	safety	3069.154
6	gpt	32	assessment	0.623751	chatbot	2639.978
7	aisi	32	innovation	0.600352	meta	2001.119
8	china	28	regulation	0.595793	gpt	1724.333
9	meta	24	ai system	0.549342	innovation	1446.921
10	claude	24	terror	0.532428	claude	1328.561
11	test	24	security	0.525495	deregulation	1277.467
12	uk	24	evaluation	0.501565	evaluation	1193.264
13	framework	20	exploitation	0.483196	legislation	1175.683
14	research	20	safeguard	0.465221	china	1173.9
15	agent ai	20	standard	0.454398	report	985.8754
16	eu	20	research	0.426112	responsible ai	968.1277
17	openai	20	weapon	0.42202	research	965.0054
18	evaluation	16	project	0.41461	agent ai	849.959
19	report	16	report	0.402107	aisi	822.6363
20	security	16	cyber security	0.399884	benchmark	770.4556

Appendix B. Network Topology