AI Safety Forecast Report*

Yeongkyun Jang, Goun Yoo, Minnseok Choi
AI Safety Policy and Strategic Cooperation, Korea AI Safety Institute

1. Overview

As AI systems become more capable and more deeply embedded in everyday life, questions about their safety are no longer confined to technical discussions. Concerns about how AI is governed, who is responsible when things go wrong, and how risks are managed at scale are now shaping global policy debates. At the same time, responsibility for AI safety is increasingly shared among governments, private companies, and specialized institutions, while geopolitical competition adds another layer of pressure. In this environment, understanding how AI safety debates are structured, rather than simply which issues attract attention, has become essential.

This report responds to that challenge by examining global discussions on AI safety through network centrality analysis of major news coverage from November 2025 to January 2026. By looking at how key concepts, institutions, and actors are connected, the analysis reveals which ideas anchor the debate, which actors bridge different policy domains, and which dynamics are likely to shape AI safety governance over time. Building on these insights, the report develops three scenario pathways that explore how regulation, industry-led frameworks, and geopolitical forces may interact in the years ahead. Together, the analysis and scenarios offer a clearer picture of where AI safety governance may be headed and how policymakers and institutions can prepare for an increasingly complex global landscape.

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

This section examines how global discussions on AI safety have evolved, using network centrality analysis of key concepts, institutions, and actors. By jointly analyzing weighted degree centrality, eigenvector centrality, and betweenness centrality, the study moves beyond simple counts of attention and highlights how influence is structured within the AI safety discourse. This approach helps identify which ideas anchor the debate, which actors connect different policy domains, and which trends are likely to shape AI safety governance over time.

2.1 Regulation as the Core Anchor of AI Safety Governance

* 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=

The analysis clearly shows that regulation has emerged as the most influential organizing concept in the AI safety network. It ranks first in both weighted degree and eigenvector centrality, indicating that regulation is not only widely discussed but also deeply embedded in connections with other influential ideas. This pattern suggests a fundamental shift in how AI safety is understood. Rather than being treated as a voluntary commitment or a purely technical concern, AI safety is increasingly framed as a matter of formal governance and public responsibility.

Regulation also plays a strong connecting role across the network. Its high betweenness centrality indicates that it links discussions on risk, safety, evaluation, and governance frameworks. In practice, this reflects growing expectations that AI risks should be addressed through enforceable rules, institutional oversight, and clearly defined accountability structures. As a result, regulation now functions as the backbone through which diverse AI safety concerns are brought together into a more coherent governance agenda.

2.2 The Increasing Convergence of Risk, Safety, and Security

Risk, safety, and security together form the conceptual core of contemporary AI safety discourse. Each of these terms ranks highly across the centrality measures, demonstrating both sustained prominence and close interconnection. Risk remains the foundational concept that motivates concern, while safety and security increasingly shape how risks are interpreted, prioritized, and managed.

This convergence reflects a broader understanding of what AI-related harm can entail. AI safety is no longer limited to technical malfunctions or model errors. Instead, it increasingly encompasses societal harm, threats to public trust, and concerns linked to national and international security. The strong presence of topics such as deepfakes, child protection, and control further illustrates this expansion, highlighting growing attention to information integrity, protection of vulnerable populations, and the need for reliable oversight of powerful AI systems.

2.3 Evaluation and Frameworks as Pathways to Institutionalization

Evaluation and frameworks have become central tools through which AI safety is translated from principle into practice. Evaluation ranks highly across all centrality measures and plays an especially important bridging role within the network. This indicates that assessment processes are now seen as essential mechanisms for turning regulatory goals into operational requirements.

These processes include impact assessments, system assessments, reporting obligations, and ongoing monitoring practices. Frameworks support this development by providing structure and coherence across different policy and technical domains. Rather than relying on isolated tests or ad hoc measures, AI safety is increasingly organized around comprehensive frameworks that align regulation, evaluation, safety standards, and reporting practices. Together, these trends suggest that AI safety governance is entering a more mature phase focused on consistency, durability, and long-term effectiveness.

2.4 The Connecting Role of AISI and Key National Actors

The betweenness centrality results highlight AI Safety Institute as the most important connecting node in the network. This position suggests that specialized AI safety institutions play a crucial role in linking national policy debates, technical expertise, and international cooperation efforts. By operating at the intersection of research, governance, and diplomacy, such institutions help translate complex technical risks into shared policy approaches.

The United States and the United Kingdom also occupy prominent bridging positions, reinforcing their influence in shaping and spreading AI safety norms. Along with the European Union, Australia, and institutions such as NIST, these actors form a dense and interconnected governance network. Through this network, regulatory ideas, evaluation practices, and safety frameworks circulate across borders, supporting gradual convergence even in a diverse and evolving global policy environment.

2.5 Growing Attention to AI Systems and Agentic AI

Although AI systems and agentic AI do not dominate the network in terms of overall frequency, their relatively high eigenvector centrality points to rising strategic importance. This suggests that attention is gradually shifting away from static or narrowly defined models toward more autonomous, interactive, and continuously operating systems.

The close association of agentic AI with concepts such as control, action plans, and cooperation reflects increasing recognition that these systems pose distinct governance challenges. Managing their risks requires not only safeguards at the point of deployment, but also sustained oversight, institutional coordination, and in some cases international cooperation. Taken together, these trends indicate that AI safety discourse is beginning to anticipate a future in which system-level and agent-driven risks become central concerns for policymakers, regulators, and AI safety institutions alike.

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

Drawing on the results of the network centrality analysis, the expert group emphasized that the most influential keywords should be understood as long term structural drivers rather than short lived signals tied to specific events. Concepts such as regulation, risk, safety, security, evaluation, and framework consistently appeared at the top of multiple centrality measures, suggesting that they form the backbone of the current AI safety discourse. The experts noted that high weighted degree reflects what is most visible in ongoing debates, while high eigenvector centrality points to deeper influence through strong connections with other important concepts. Keywords with high betweenness centrality were seen as especially significant, as they connect otherwise separate policy, technical, and geopolitical discussions and help explain how shifts in one area can reshape the broader governance landscape.

Based on this shared understanding, the expert group organized the key driving factors into three broad and closely linked thematic areas. The first centers on policy and governance and includes regulation, evaluation, frameworks, and safety, reflecting the growing role of formal oversight and institutionalized responsibility in AI safety. The second focuses on industry and technical systems, bringing together concepts such as AI systems, agentic AI, testing, and reporting, which capture the changing risk profile of more autonomous and interactive technologies. The third relates to geopolitics and international coordination, highlighting security concerns and the influence of major actors such as the United States, the United Kingdom, the European Union, and China. The experts stressed that these areas cannot be treated in isolation, as developments in one consistently shape and constrain developments in the others.

Using these thematic groupings as a foundation, the expert group developed a scenario framework guided by the principle of structural stability. Highly central concepts were treated as forces likely to remain influential even as policies, technologies, and political conditions evolve. Scenarios were constructed by exploring how different combinations of these driving factors might reinforce one another or come into tension over time. Rather than aiming to predict a single future, the scenarios are intended to capture dominant directions that may overlap or unfold in parallel. This approach allows policymakers and institutions to think more clearly about uncertainty, identify meaningful leverage points, and prepare adaptive strategies in a global AI environment shaped by rapid technological change and growing geopolitical complexity.

4. Scenario Forecasts Based on Key Driving Factors

The network analysis and expert discussions point to a small number of forces that are likely to shape the future of AI safety over an extended period. Regulation, safety, evaluation, industry frameworks, and geopolitical dynamics are not short-term trends, but structural elements that will continue to interact as AI systems become more capable and more widely deployed. The three scenarios below describe different ways these forces may combine. They should not be read as predictions or as mutually exclusive outcomes. Instead, they offer plausible pathways that highlight where influence could concentrate, how responsibilities might be distributed, and what kinds of trade-offs policymakers and institutions may face as AI safety governance evolves.

4.1 Scenario A. Governance-Led Consolidation around Regulation and Safety

In this scenario, governments respond to growing concerns about AI-related harm by firmly placing safety within the domain of public governance. As AI systems become more capable and more deeply embedded in society, policymakers increasingly view voluntary commitments and informal guidance as insufficient. Regulation and safety therefore become the primary organizing principles of AI governance. Legal requirements expand beyond high-level rules to include concrete obligations such as risk assessments, documentation, reporting, and ongoing monitoring. Public authorities strengthen their oversight capacity, and specialized institutions play a larger role in translating technical risks into enforceable standards. The European Union continues to shape global expectations through the implementation of the EU AI Act, while other countries adopt comparable approaches or align parts of their systems to remain interoperable.

At the same time, major powers follow distinct regulatory paths. The United States combines formal rules with flexible governance tools and close engagement with industry, while China advances a state-centered model that emphasizes security and administrative control. Despite these differences, this scenario is characterized by a shared understanding that AI safety is a matter of public responsibility requiring durable institutional foundations. This consolidation brings greater clarity and accountability, but it also introduces new challenges. Regulatory systems may struggle to keep pace with rapid technological change, and compliance costs could weigh more heavily on smaller developers. The central tension in this scenario lies in maintaining strong safeguards without undermining innovation or adaptability.

4.2 Scenario B. Industry-Led Systematization through Frameworks and Evaluation

In this scenario, much of the practical work of AI safety continues to be driven by frontier AI companies. As these firms confront increasingly complex risks, they invest heavily in internal safety processes such as red teaming, capability testing, continuous evaluation, and incident response. Over time, these practices become more formalized and evolve into comprehensive safety frameworks. Because these companies operate at the cutting edge of AI development, their tools and methodologies often advance faster than public regulation. Regulators and policymakers, facing capacity constraints and fast-moving technologies, increasingly rely on industry-developed evaluation methods to operationalize safety goals. This creates a hybrid governance environment in which public authorities define broad expectations, while private frameworks provide the technical detail needed to implement them.

The main strength of this scenario is speed and adaptability. Industry-led systems can respond quickly to emerging risks, particularly those associated with agentic AI and highly interactive systems. However, this model also raises important questions. Concentrating influence in a small number of firms may limit transparency and create uneven power dynamics, while smaller developers may find it difficult to meet evolving expectations. The central challenge in this scenario is ensuring that industry leadership in safety remains accountable, open, and aligned with public interest, rather than becoming a substitute for public oversight.

4.3 Scenario C. Geopolitical Fragmentation and Security-First AI Safety

In this scenario, AI safety becomes increasingly shaped by geopolitical competition and national security concerns. As strategic rivalry intensifies, especially between the United States and China, governments begin to treat AI safety frameworks as instruments of economic and security policy rather than as shared global goods. Countries develop their own certification systems, testing requirements, and safety standards, and mutual recognition across borders weakens. Safety expectations become closely tied to export controls, alliance structures, and defense considerations. Information sharing on AI incidents, vulnerabilities, and evaluation results becomes more limited, even among partners, as trust erodes and security concerns take precedence.

This fragmentation strengthens domestic control over sensitive technologies but comes at a cost. Without shared

frameworks or reliable channels for cooperation, blind spots emerge, and collective preparedness for systemic AI risks declines. As AI systems grow more autonomous and interconnected, these gaps become harder to manage. The core tension in this scenario lies between control and coordination. While governments may gain greater confidence in their ability to manage risks at home, the absence of meaningful international cooperation increases the likelihood of uneven safety outcomes and delayed responses to emerging global threats.

5. Strategic Directions

Taken together, the analysis and scenarios in this report point to a clear and pressing challenge. AI safety governance must remain coherent in an environment where regulatory systems, technological capabilities, and geopolitical pressures are evolving in different directions at the same time. As rules multiply and safety frameworks develop across jurisdictions and industries, the central risk is not inaction but misalignment. A core strategic priority is therefore to establish a shared baseline for AI safety, including common understandings of risk, evaluation practices, and transparency expectations. Such a baseline would allow different governance models to remain interoperable even when full harmonization is neither realistic nor desirable.

A second strategic priority is to strengthen the practical foundations that support AI safety across the broader ecosystem. As evaluation and framework-based approaches become more central, safety capabilities should not be concentrated in the hands of a small number of leading firms. Investment in shared testing infrastructure, accessible assessment tools, and credible third-party assurance mechanisms can help distribute safety capacity more widely. This is especially important as AI systems become more autonomous and interconnected, making risks harder to detect through one-time reviews or static compliance requirements.

Finally, the scenarios underscore the importance of sustaining cooperation under conditions of growing geopolitical competition. Even when trust is limited and political alignment is difficult, fully isolated approaches increase the likelihood of blind spots and delayed responses to emerging risks. Maintaining focused and reliable channels for information sharing, such as incident reporting, vulnerability analysis, and joint research on new risk profiles, can support early warning and collective preparedness. In this environment, regional partnerships, specialized AI safety institutions, and structured public private collaboration offer practical ways to manage shared risks while navigating an increasingly contested global landscape.

Appendix A. Results of Network Centrality Analysis

Rank	Weighted-degree centrality		Eigenvector centrality		Betweeness centrality	
1	regulation	51	regulation	1	aisi	7207.692
2	risk	31	risk	0.955296	us	6350.144
3	security	30	security	0.910098	uk	3691.437
4	evaluation	44	evaluation	0.83056	safety	2832.265
5	framework	26	framework	0.726081	evaluation	2688.19
6	aisi	48	aisi	0.69785	eu	2642.191
7	safety	42	safety	0.697008	regulation	2483.825
8	report	17	report	0.665777	security	2398.253
9	test	11	test	0.62719	openai	2236.972
10	experiment	14	experiment	0.617815	risk	2085.94
11	deepfake	24	deepfake	0.583044	content	1728.885
12	child	20	child	0.571744	deepfake	1537.114
13	control	21	control	0.517107	report	1427.718
14	system	13	system	0.472798	gpt	1409.56
15	trend	9	trend	0.446394	control	1251.163
16	governance	8	governance	0.444968	act	1129.172
17	action plan	10	action plan	0.442375	australia	1104.474
18	cooperation	10	cooperation	0.439395	election	1060
19	ai system	7	ai system	0.4379	nist	1048.801
20	agent ai	16	agent ai	0.432165	guideline	983.9736

Appendix B. Network Topology