AI Safety Forecast Report*

Goun Yoo, Yeongkyun Jang, Minnseok Choi

AI Safety Policy and Strategic Cooperation, Korea AI Safety Institute

1. Overview

This AI Safety Forecast Report tracks how global conversations on AI safety are shifting as technology advances and policy priorities change. The analysis draws on major news coverage from February to April 2026, a period in which AI safety discussions have grown more complex, interconnected, and operationally focused. Previous editions identified risk (August 2025), the United States (November 2025), and regulation (February 2026) as the leading keywords, each ranking first in at least two of the three centrality metrics. The present analysis builds on that trajectory: governance now ranks first across all three measures for the first time, marking it as the new structural center of global AI safety discourse.

The report goes beyond asking which topics appear most frequently. Through network centrality analysis, it maps how key concepts, actors, and issue areas relate to one another, identifying how different elements of AI safety are linked and how influence is distributed across the broader conversation. The goal is to move past surface-level trends and into the structural dynamics actually shaping AI safety today.

Building on this structural perspective, the report outlines three forward-looking governance pathways: the strengthening of evaluation-centered governance, the growing importance of managing risks from increasingly autonomous AI systems, and the emergence of more distributed forms of coordination and accountability across the AI safety ecosystem. Each scenario highlights a different dimension of governance, but all three point toward the same underlying shift from principle-based and actor-centered approaches toward more operational, system-oriented, and behavior-aware models. Read as a set, they offer a grounded view of how AI safety governance may evolve in an environment that is growing more complex, interconnected, and uncertain.

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

This section examines how global discussions on AI safety are currently structured, based on network centrality analysis. The focus is not on frequency alone, but on how key concepts connect to one another and how influence flows across

* 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 broader discourse. That structural lens helps explain not just which issues matter, but also how they are related to one another and how AI safety governance is evolving over time.

2.1 Governance as the Central Anchor of AI Safety

The most notable finding from the analysis is the strong centrality of governance, which ranks first across all major metrics. AI safety is no longer viewed as a narrow technical concern; it is increasingly understood as a broader challenge that requires coordination across institutions, policies, and stakeholders. The discussion has shifted toward how AI systems should be governed in a consistent and accountable way. This position is unprecedented in the history of this report. In previous editions, risk (August 2025), the United States (November 2025), and regulation (February 2026) each ranked first in at least two of the three metrics, but none achieved the top position across all three simultaneously. The shift suggests that the discourse has moved beyond identifying the core concern, namely risk; determining who should lead, namely national actors; or deciding which instrument should be used, namely regulation. Instead, it now centers on a more encompassing question: how AI safety should be organized and sustained as a system. Governance now functions not merely as one element among others, but as the overarching structure that connects regulation, policy, and accountability into a coherent framework.

2.2 Evaluation as the Bridge Between Principle and Practice

Within this governance-centered structure, evaluation plays a crucial role in linking high-level principles to real-world practice. Its prominence across centrality measures reflects a growing demand to assess whether AI systems are actually safe and reliable, not just in principle, but under real-world conditions. Evaluation here encompasses benchmarking, testing, and continuous monitoring: processes that make safety concrete and measurable. Through these mechanisms, stakeholders can move past general commitments and toward verifiable outcomes. AI safety is becoming more practical and structured as a result, with evaluation doing the work of translating governance goals into something that can actually be checked.

2.3 Convergence of Risk, Safety, and Security

The analysis also reveals a strong connection between risk, safety, and security, which together form the core of AI safety discourse. Risk names the potential harms; safety is the goal of reducing them; security widens the frame to include broader societal and strategic concerns. The growing prominence of concepts such as cybersecurity and deepfake reflects how AI-related risks are becoming more complex. Concerns now extend beyond technical errors to include malicious use, information manipulation, and wider social impacts, a sign that AI safety is maturing into a more comprehensive framework that addresses both technical vulnerabilities and external threats.

2.4 Hybrid Governance Shaped by Industry and Geopolitics

The network highlights the important roles played by both industry and governments. Companies such as Anthropic and OpenAI are at the forefront of developing practical safety approaches, particularly in evaluation and testing. Their technical proximity to the systems lets them iterate quickly and shape emerging best practices before formal standards catch up. At the same time, governments, regional institutions, and specialized organizations continue to shape the broader direction of AI governance. Actors such as the United States, China, and the European Union bring different strategic interests, regulatory traditions, and approaches to technological oversight. The result is a hybrid governance environment in which public and private actors jointly determine how AI safety evolves, while geopolitical factors shape the boundaries of cooperation and competition.

2.5 Emergence of Agentic AI and Expanding Risk Boundaries

The growing prominence of terms such as AI agent and agent marks a shift toward more autonomous and interactive AI systems. These systems make decisions and take actions over time, introducing risks that are qualitatively different from those tied to earlier, more static models. As these systems spread, the focus of AI safety is expanding accordingly. A snapshot evaluation at deployment is no longer sufficient; how a system behaves over longer periods and in changing environments matters just as much. Continuous monitoring and adaptive governance move from nice-to-have to necessary.

Beyond the primary driving factors discussed above, several keywords in the lower ranks of the centrality analysis deserve attention as potential early signals. Behavior (betweenness rank 17) and benchmark (weighted-degree and eigenvector rank 20) hover at the periphery of the top-20 across different metrics, signaling relevance that has not yet consolidated into primary driving-factor status. Cyber (betweenness rank 19) similarly stays outside the top tier, though it may gain ground as AI-enabled threats grow in scale. None of these terms qualifies as a key driving factor yet, but their consistent appearance across multiple centrality measures makes them worth tracking in the quarters ahead.

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

Drawing on the network centrality analysis, the expert group interpreted the current AI safety discourse as entering a more applied oversight and system-oriented phase. The discussions moved away from treating highly ranked keywords such as governance, evaluation, safety, risk, and security as isolated topics, emphasizing instead how these concepts work together as interconnected components of a single governance system. Governance, in this reading, is not simply regulation or institutional authority; it is the broader framework through which safety expectations, accountability structures, technical oversight, and implementation mechanisms are coordinated across the AI ecosystem. Evaluation, meanwhile, was identified as the critical operational mechanism that turns broad safety principles into measurable practices: benchmarking, testing, monitoring, auditing, and verification.

The expert group also noted that the structure of AI safety discourse appears to be shifting away from debates centered primarily on leadership and institutional competition toward questions of operational governance. Earlier discussions in

AI safety often focused on which actors, such as governments, frontier AI companies, or international institutions, would shape global norms and standards. The current analysis suggests a different preoccupation: how governance can actually work as AI systems grow more capable, autonomous, and deeply embedded in social, economic, and security infrastructures. In that light, concepts such as monitoring, transparency, benchmarking, and behavior read as signals of growing attention to implementation, oversight, and long-term system conduct.

Based on these interpretations, the expert group identified three broader governance transitions that may shape the future of AI safety. The first is the operationalization of governance through evaluation-centered oversight: safety becoming measurable, testable, and enforceable through benchmarking, continuous monitoring, and verification. The second is the challenge of governing agentic and autonomous AI systems, where the safety focus shifts from isolated outputs to long-term behavior, adaptive decision-making, and evolving operational risks. The third is the growing need for distributed coordination across a fragmented AI safety ecosystem, as governments, frontier AI companies, AI safety institutes, independent evaluators, and civil society actors are called to share responsibility for oversight, accountability, and risk management in ways they have not before.

These governance transitions do not represent mutually exclusive futures. They reflect overlapping dynamics that are increasingly entangled within the AI safety landscape. What the discussions collectively suggest is a move away from a primarily principle-based and actor-centered framework toward a more operational, behavior-aware, and system-oriented model of governance.

4. Scenario Forecasts Based on Key Driving Factors

The network centrality analysis and expert group discussions suggest that the next phase of AI safety will not be defined primarily by who leads governance, whether governments, companies, or major powers. Previous reports have already explored scenarios centered on regulatory institutionalization, industry-led safety frameworks, and geopolitical fragmentation. Each of those earlier scenarios was organized around the question of which actor or force would dominate. In this report, the focus shifts from who leads to how governance operates. The fact that governance and evaluation consistently rank first and second across all three centrality measures indicates that the discourse has entered a more operational phase, where the central question is no longer whether AI safety should be governed, but through what mechanisms governance can be made measurable, adaptive, and enforceable in practice.

Governance and evaluation consistently appear among the most central concepts across all three centrality measures. This pattern points to a phase where the question is no longer whether AI safety should be governed, but through what mechanisms governance can be made measurable, adaptive, and enforceable in practice. At the same time, the growing prominence of concepts such as behavior, monitoring, and agentic AI suggests that attention is expanding beyond static models and isolated outputs toward long-term system conduct and evolving operational risks. Together, these shifts point to a broader transition in AI safety governance, from debates centered on principles and leadership structures toward the development of durable mechanisms for implementation, oversight, and accountability.

The three scenarios below describe different pathways for AI safety governance. Scenario A traces the rise of evaluation-centered governance, where safety principles are translated into measurable and repeatable procedures. Scenario B examines behavioral risk governance for agentic AI, where the unit of analysis shifts from static outputs to long-term system behavior. Scenario C looks at distributed accountability across the AI safety ecosystem, as responsibility spreads among governments, AI safety institutes, frontier AI companies, independent evaluators, and users. The three are not mutually exclusive; they represent overlapping pathways through which AI safety governance may develop.

4.1 Scenario A. Evaluation-Centered Operational Governance

The key driving factors shaping this scenario are governance, evaluation, benchmark, transparency, monitoring, and safety. The central question it addresses is operationalization: how safety can actually be verified, measured, and demonstrated. This is a future in which AI safety governance moves beyond broad rules or institutional frameworks and organizes itself around concrete evaluation and verification processes. As AI systems are deployed across more sectors, general commitments to safety stop being enough. Policymakers, developers, and oversight bodies face growing pressure to show whether AI systems are genuinely safe, reliable, and accountable.

Under these conditions, benchmarking, pre-deployment testing, documentation, continuous monitoring, and post-deployment verification become core elements of AI safety governance. Evaluation moves from a secondary technical process that supports regulation to the operational mechanism through which governance is made actionable. Governments and specialized institutions define evaluation expectations, companies integrate them into development and deployment pipelines, and independent actors help verify whether safety claims are credible. Through this process, AI safety shifts from a principle-based agenda to a more measurable, evidence-based governance system.

The strength of this scenario lies in its practicality. Evaluation-centered governance can reduce ambiguity, support comparability across systems, and provide a shared basis for accountability. It can also help bridge the gap between policy objectives and technical implementation. But this pathway carries real risks. If evaluation methods become too rigid, they may miss emerging risks or fail to reflect real-world system behavior. If benchmarks come to dominate the safety conversation, developers may optimize for test scores at the expense of actual robustness. The core design challenge is to build evaluation systems standardized enough to support accountability, yet flexible enough to keep pace with changing AI capabilities and risk profiles.

4.2 Scenario B. Behavioral Risk Governance for Agentic AI

The key driving factors here are AI agent, behavior, risk, safety, evaluation, and monitoring. Where Scenario A focuses on measuring outputs, this one asks a different question: how do we observe, evaluate, and manage the evolving behavior of increasingly autonomous AI systems over time? It describes a future in which the main challenge of AI safety shifts from assessing model outputs at a single point in time to governing the behavior of autonomous and interactive systems

as they operate. As AI agents grow more capable of planning, using tools, interacting with external environments, and carrying out multi-step tasks, risks increasingly arise from chains of actions, not individual responses.

Under this scenario, AI safety governance expands from static testing to behavior-oriented oversight. The question is no longer simply whether a model produces safe outputs in a controlled setting, but how an AI system behaves while pursuing goals, responding to shifting contexts, interpreting constraints, and interacting with users or other systems. That demands new forms of behavioral evaluation: long-term monitoring, scenario-based stress testing, intervention mechanisms, and post-deployment feedback loops. Safety assessment becomes more dynamic, tracking whether systems stay within intended boundaries as their operating environments change.

The significance of this scenario is that it expands the unit of AI safety analysis from the model to the system. It directly reflects the growing importance of agentic AI and the expansion of risk boundaries. However, behavioral risk governance also creates major challenges. Long-term system behavior is difficult to measure, compare, and audit. Developers and regulators may disagree over what counts as acceptable autonomy, meaningful human oversight, or sufficient intervention capacity. There are also practical burdens associated with continuous monitoring and post-deployment adjustment. The fundamental challenge is therefore to build governance mechanisms that can observe and manage evolving AI behavior without making AI systems unusable, overly restricted, or impossible to innovate.

4.3 Scenario C. Fragmented Governance and Distributed Accountability

The key driving factors here are governance, transparency, evaluation, security, AISI, Anthropic, OpenAI, and policy. The institutional challenge this scenario addresses is coordination: how to distribute accountability across an AI safety ecosystem that is fragmenting and drawing in more actors. It describes a future in which AI safety governance is not organized around a single dominant authority, but held together by the interaction of governments, frontier AI companies, specialized institutions, independent evaluators, and civil society actors operating across different jurisdictions and governance systems.

As AI systems embed more deeply into economic, social, and security infrastructures, responsibility for safety spreads across a wider set of actors. Governments set baseline obligations, reporting requirements, and public accountability principles, but they often lack direct access to frontier model development and deployment processes. Frontier AI companies have the technical capacity to build and operate advanced systems, but their internal safety claims need external validation and public credibility. AI Safety Institutes and other specialized organizations contribute evaluation, risk analysis, technical standards, and policy coordination, while independent researchers, auditors, and civil society organizations help surface failures, misuse, and emerging harms in real-world environments. AI safety governance, in this picture, becomes a system of distributed but interdependent oversight.

The significance of this scenario lies in its recognition that no single actor can fully govern advanced AI systems alone. Distributed governance can combine different forms of expertise, oversight capacity, and institutional legitimacy,

reducing overreliance on either state regulation or private self-governance. However, this ecosystem also introduces major coordination challenges. Fragmented accountability may create gaps between developers, deployers, regulators, and evaluators, especially when systems operate across jurisdictions or when security and commercial pressures limit transparency. Conflicts may also emerge over which institutions have the authority to define acceptable risk, evaluate compliance, or intervene in high-risk deployments. The core challenge, then, is not only distributing responsibility but maintaining coordination, transparency, and enforceability across a fragmented governance environment. Clear reporting structures, audit mechanisms, escalation procedures, and trusted channels for cross-institutional information sharing will become essential for sustaining accountability as AI governance grows more complex and contested.

5. Strategic Directions

The analysis and scenarios in this report point to a structural shift in how AI safety must be governed. As governance, evaluation, risk, and system behavior become more tightly interconnected, the defining challenge is no longer articulating high-level safety principles, but operationalizing them in ways that hold up as AI systems evolve. The first strategic priority is therefore to move from principle-based governance toward evaluation-centered operational systems. That means building out benchmarking, continuous monitoring, and post-deployment verification so that safety can be demonstrated and enforced on the ground, not merely declared through compliance. Governance and evaluation have ranked first and second across all three centrality measures this quarter. This pattern reinforces the broader direction of the discourse by showing that evaluation has become the operational core through which governance earns its credibility.

A second priority is to adapt AI safety governance to the rise of agentic and autonomous systems. As the scenarios illustrate, risks increasingly come from chains of actions and long-term system behavior, not isolated outputs. Governance therefore needs to shift from static model evaluation toward behavior-oriented oversight, including scenario-based testing, real-time monitoring, and intervention mechanisms. Developing the capacity to observe and manage evolving system behavior will be critical to keeping safety frameworks relevant as AI systems grow more interactive, adaptive, and capable of independent decision-making. The fact that behavior now registers in the centrality analysis is itself a signal: governance frameworks need to be built to handle dynamic and evolving system conduct.

Finally, the scenarios point to the need for clearer structures of distributed accountability across the AI safety ecosystem. As responsibility spreads among governments, frontier AI companies, AI safety institutes, independent evaluators, and users, the risk is not the absence of oversight but its fragmentation. The priority is to make roles, obligations, and information flows more explicit and enforceable by strengthening reporting mechanisms, auditability, and feedback loops so that safety-relevant information can move efficiently across actors and support timely intervention. With security risks and geopolitical pressures mounting, that kind of structured coordination will be essential to maintaining both accountability and adaptability.

Appendix A. Results of Network Centrality Analysis

Rank	Weighted-degree centrality		Eigenvector centrality		Betweenness centrality	
1	governance	116	governance	1	governance	8513.843
2	evaluation	64	evaluation	0.701664	evaluation	3248.435
3	us	64	safety	0.691662	safety	3125.066
4	safety	60	risk	0.650507	us	2866.510
5	anthropic	56	us	0.649602	security	2842.442
6	security	52	anthropic	0.626987	risk	2696.379
7	risk	52	security	0.544718	anthropic	2647.038
8	aisi	48	regulation	0.536114	aisi	2451.496
9	regulation	40	aisi	0.509593	regulation	1358.784
10	china	36	policy	0.475460	ai agent	1349.602
11	policy	36	china	0.457169	uk	1203.820
12	deepfake	32	cybersecurity	0.4280377	openai	1144.103
13	eu	32	transparency	0.4197482	deepfake	1135.361
14	ai agent	28	openai	0.4051679	china	1100.184
15	uk	28	ai agent	0.3855494	cybersecurity	1093.762
16	openai	28	deepfake	0.3439814	eu	1054.221
17	cybersecurity	28	agent	0.3339992	behavior	987.3951
18	agent	28	standard	0.3307404	ai safety	968
19	transparency	28	uk	0.3186489	cyber	968
20	benchmark	24	benchmark	0.3130795	agent	884.9541

** Note: This is the fourth edition of the network centrality analysis. Governance is the first keyword to rank first across all three centrality metrics simultaneously. Results from previous editions are available on the Korea AISI website.*

Appendix B. Network Topology