# 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 examines how global discussions on AI safety evolved through network centrality analysis of major news coverage from May to July 2026. During this period, rapid advances in frontier AI models were accompanied by growing efforts to strengthen AI safety and security governance. The analysis identifies a structural shift at the center of the network. For the first time across the five editions of this report, AISI ranks first across all three centrality measures (weighted-degree, eigenvector, and betweenness centrality), while Anthropic ranks second on all three measures. These rankings should be interpreted within the scope of the news sources monitored for this report, which reflect the Korea AI Safety Institute's institutional focus and monitoring priorities rather than a representative sample of all global AI safety news. Within this monitoring scope, the progression of leading keywords from risk (August 2025), to the United States (November 2025), regulation (February 2026), governance (May 2026), and AISI (August 2026) suggests a shift toward greater attention to the institutions responsible for implementing AI safety governance.

Beyond this shift, cybersecurity and AI evaluation rank immediately behind the two leading keywords, while AI agent, MOU, and Mythos appear in the top rankings for the first time. These developments reflect the growing importance of agentic AI risks, expanding bilateral cooperation among AI Safety Institutes, and the policy impact of Anthropic's temporary suspension of its Mythos-tier models in response to U.S. export controls in June 2026. Based on these patterns, the research team develops three scenarios. The first examines the consolidation of an international AI Safety Institute network as a foundation for global AI safety governance. The second explores increasing competition among frontier AI companies to secure agentic AI systems. The third considers the growing convergence of AI safety with cybersecurity, CBRN, and autonomous weapons. Taken together, these scenarios suggest that AI safety governance is entering a phase in which institutional roles and responsibilities have become as important as governance mechanisms themselves.

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

This section examines how AI safety issues are structured within the international news sources monitored for this report

* 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=

using network centrality analysis. Rather than focusing on frequency alone, it considers the structural prominence of key concepts within the monitored network and how their relative positions have changed over time.

### 2.1 AI Safety Institutes Move to the Center of the Network

For the first time in this report series, AISI ranks first across all three centrality measures, while Anthropic ranks second on each measure. Within the monitoring scope of this report, this finding indicates increased prominence of institutions involved in implementing AI safety policies and evaluation activities. Given the institutional focus of the monitoring process, however, AISI's leading position should not be interpreted as evidence that it is the most central actor across the entirety of global AI safety discourse.

The prominence of AISI coincides with a series of bilateral cooperation agreements among national AI Safety Institutes. In July 2026, the Korean and Canadian AI Safety Institutes signed a memorandum of understanding to strengthen cooperation on AI risk information sharing, evaluation methodologies, and policy frameworks. This followed the Australia–Canada AISI agreement signed earlier in the year alongside the launch of Australia's AI Safety Institute. Since the first international meeting of AI Safety Institutes was held in San Francisco in November 2024, cooperation among these institutions has continued to expand and has become an increasingly important element of global AI safety governance.

The emergence of MOU as a prominent keyword (14th in weighted-degree centrality and 9th in eigenvector centrality) further reflects this trend. Rather than relying on a single international agreement, countries are strengthening cooperation through bilateral memoranda of understanding. Together with the central position of AISI, these developments suggest that AI safety governance is evolving through closer collaboration among specialized national institutions while maintaining a distributed governance structure rather than a single centralized authority.

### 2.2 Growing Importance of Cybersecurity in AI Evaluation

Cybersecurity ranks third in weighted-degree centrality and fourth in eigenvector centrality, while AI evaluation ranks fourth and fifth, respectively. Their similarly high rankings indicate that both topics have become prominent within the broader AI safety discourse. Developments during the reporting period further suggest increasing overlap between cybersecurity and AI evaluation in policy and practice.

This trend is reflected in several developments during the reporting period. In July 2026, the European Commission introduced an Action Plan on Cybersecurity and Artificial Intelligence, developed in cooperation with ENISA. The plan explicitly connects compliance with the AI Act to existing cybersecurity legislation, including the NIS2 Directive and the Cyber Resilience Act. During the same month, reports that an unreleased frontier AI model breached its sandbox environment during an internal security test highlighted the growing importance of cybersecurity in evaluating advanced AI systems.

The increasing integration of cybersecurity and AI evaluation is also reflected in recently published safety benchmarks for AI agents. These evaluations found that none of the production-grade AI agents tested completed more than 40 percent of assigned tasks while complying with all safety requirements. The findings remain significant even as the EU AI Act's Annex III high-risk AI system obligations, originally set to take effect on August 2, 2026, were postponed to December 2, 2027 following the EU's "AI Omnibus" agreement (in force July 27, 2026): the compliance gap the benchmarks reveal will still need to be closed before the new deadline. The continued prominence of red teaming and benchmark among the leading keywords further suggests that structured adversarial evaluation is becoming a standard component of AI system development and deployment.

### 2.3 From Principles to Practice: Governance, Transparency, and International Cooperation

Governance and transparency rank third and fourth in betweenness centrality, respectively. Their high betweenness centrality suggests that each occupies an important bridging position within the broader AI safety network. Their prominence, alongside developments during the reporting period, points to growing attention to the implementation and accountability of AI governance.

Several developments during the reporting period illustrate this trend. Illinois became the first U.S. state to enact legislation requiring annual independent audits of frontier AI models, while the European Union continued to expand guidance for implementing the AI Act's risk classification framework. Together, these developments indicate a broader shift from announcing AI safety commitments to demonstrating their implementation through regulatory and evaluation mechanisms.

In mid-2026, an industry AI safety index evaluated nine leading AI companies across a broad set of transparency and risk assessment indicators. This reflects growing expectations that AI governance should be supported by independently verifiable evidence rather than policy commitments alone.

### 2.4 Agentic AI and the Widening Capability-Safety Gap

AI agent ranks eighth in both weighted-degree and eigenvector centrality and ninth in betweenness centrality. Related terms including AI model, GPT, LLM, Frontier AI, OpenAI, and Google DeepMind also appear prominently in the network. The prominence of AI agent, together with the presence of related model- and developer-level terms, suggests growing attention to agentic AI systems within the broader AI safety discourse.

Reports published during the reporting period also point to rapid growth in the use of AI agents and agentic web browsers. This growth has been concentrated among a relatively small number of organizations, increasing both their influence in the AI ecosystem and their responsibility for ensuring the safe deployment of these systems.

At the same time, recent evaluations have highlighted differences between agent performance in controlled testing environments and real-world deployment. A widely cited international AI safety report found that frontier AI models often demonstrate safer behavior during evaluations than after deployment. These findings raise concerns that existing evaluation methods may encourage systems to optimize for benchmark performance rather than consistently safe behavior in operational settings.

The continued prominence of benchmark, red teaming, and AI agent among the leading keywords suggests that improving the reliability of AI agent evaluations has become an important priority for both AI developers and evaluators.

### 2.5 The Mythos Episode and the Return of National-Security Framing

Mythos appears among the leading keywords for the first time in this report, ranking tenth in weighted-degree centrality and eleventh in eigenvector centrality. This reflects Anthropic's release of Claude Mythos 5 and Claude Fable 5 in June 2026, followed by a temporary suspension of service to comply with export controls imposed by the U.S. Department of Commerce. Service resumed on July 1, 2026, after the restrictions were lifted. These developments illustrate how advances in frontier AI can quickly become associated with national security considerations. A similar trend is reflected in the White House Executive Order issued on June 2, 2026, which introduced a voluntary pre-release review framework for advanced AI models from cybersecurity and national security perspectives.

This emphasis on national security extends beyond individual companies and is particularly evident in the area of CBRN risk. A representative example is the United Kingdom's AI Security Institute. In February 2025, the former AI Safety Institute was renamed the AI Security Institute, with its policy priorities shifting toward major security risks, including chemical and biological weapons, rather than broader AI safety issues such as bias or misinformation. Since then, the institute has continued to play a leading role in CBRN evaluations through its Frontier Trends Report and its partnership with the Defence Science and Technology Laboratory (DSTL). It has also expanded evaluation cooperation with members of the international AI Safety Institute network, including the United States and Singapore.

These institutional developments have been accompanied by practical policy measures. Through its Biological Security Strategy implementation report, the UK government disclosed the operation of Biothreats Radar, a real-time biological threat surveillance system used across fourteen government departments. The system was also used during the 2026 hantavirus outbreak response. At the international level, the OECD continued work on recommendations for responsible innovation in synthetic biology to strengthen biosafety and biosecurity governance. In May 2026, OpenAI also introduced a governance framework covering cybersecurity, CBRN, harmful manipulation, and loss-of-control risks. Together, these developments indicate that CBRN-related AI governance is becoming more firmly integrated into national security policy, operational preparedness, and corporate governance.

Security-related topics such as Deepfake (betweenness rank 10) and Autonomous Weapons (betweenness rank 20) also appear within the top 20. Developments during the reporting period further indicate continued attention to these issues in international AI governance discussions. During the reporting period, the United Nations held its first dedicated global dialogue on AI governance, with discussions covering non-consensual deepfakes and lethal autonomous weapons. The

G7 also continued discussions on international approaches to military AI, although no binding agreement has yet been reached. Overall, these developments suggest that AI safety discussions are increasingly incorporating national security concerns, particularly in areas such as CBRN, where institutional cooperation and governance frameworks have continued to develop.

### 2.6 Emerging Signals

A low ranking alone does not make a keyword an emerging signal; the terms discussed below are flagged because they appear consistently across multiple centrality measures or reflect newly visible actors, distinguishing them from established topics that simply rank lower.

Beyond the primary driving factors discussed above, several lower-ranked keywords deserve attention as potential early signals of emerging developments in AI safety. Frontier AI (weighted-degree rank 17) remains outside the top tier but may indicate growing attention to the governance and evaluation challenges associated with increasingly capable AI models. Resilience (eigenvector rank 19) and International cooperation (eigenvector rank 20) similarly appear at the periphery of the network, suggesting that institutional preparedness and cross-border coordination may become more prominent as AI safety governance continues to develop.

Other emerging signals point to changes in the actors involved in AI safety discussions. G7 (betweenness rank 18) and Google DeepMind (betweenness rank 19) also appear at the lower end of the top 20, indicating that international actors and frontier AI developers remain visible, though not yet central, within the network. None of these terms yet occupies a sufficiently central position to qualify as a primary driving factor. Their presence within the top 20, however, makes them worth tracking in future reporting periods as potential indicators of issues and actors that may gain greater prominence.

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

Based on the results of the network centrality analysis, the expert group interpreted the major keywords (AISI, Anthropic, Cybersecurity, AI evaluation, OpenAI, Governance, AI agent, Transparency, Mythos, MOU, National security, CBRN, UK, etc.) from three complementary perspectives, following the analytical approach used in previous editions of this report. Keywords with high weighted-degree centrality were interpreted as reflecting issues with strong current influence across the network. Keywords with high eigenvector centrality were identified as influential topics connected to other major issues and therefore likely to shape future developments. Keywords with high betweenness centrality were interpreted as connecting different areas of the AI safety discourse and facilitating interactions among otherwise separate topics.

Based on these findings, the expert group organized the keywords into three thematic clusters. The first is an institutions and governance cluster, including AISI, Governance, Transparency, MOU, International cooperation, and AI evaluation.

The second is an industry and agentic AI cluster, consisting of Anthropic, OpenAI, Google DeepMind, AI agent, AI model, GPT, LLM, Frontier AI, Benchmark, Red teaming, and Mythos. The third is a security and geopolitics cluster, including Cybersecurity, National security, CBRN, Security, Autonomous weapons, Deepfake, G7, and the UK.

Drawing on these three thematic clusters, the expert group developed three scenarios using the same analytical framework applied throughout this report series. Scenario A focuses on institutions and governance, Scenario B examines developments led by industry and agentic AI technologies, and Scenario C considers security and geopolitical dynamics. The following sections present the driving factors, future developments, and key implications associated with each scenario.

## 4. Scenario Forecasts Based on Key Driving Factors

The previous edition anticipated a growing emphasis on governance and AI evaluation. During the current reporting period, bilateral cooperation among AI Safety Institutes expanded, AI evaluation became more closely integrated with cybersecurity, and national security considerations became increasingly prominent. These developments suggest that the key directions identified in the previous edition have continued to evolve.

The network centrality analysis, interpreted within the monitoring scope of this report, together with expert group discussions, suggests a possible shift in the focus of AI safety governance. Previous editions of this report examined how AI safety governance could evolve through regulatory institutionalization, industry-led governance, and geopolitical competition. The May 2026 edition further emphasized the growing importance of governance and evaluation as the key mechanisms supporting AI safety. In this reporting period, the findings indicate another important shift. Within the monitored discourse, attention appears to be shifting from governance mechanisms toward the institutions responsible for implementing AI safety governance.

This shift is reflected in the central positions of AISI and Anthropic, which rank first and second across all three centrality measures. Unlike previous editions, these two institutions consistently occupy the most influential positions in the network. Their prominence suggests that greater attention is being given to the organizations responsible for conducting AI evaluations, establishing technical standards, and supporting international cooperation. At the same time, the growing prominence of MOU, Mythos, CBRN, and National security indicates that bilateral cooperation, export controls, and national security considerations are becoming increasingly important in determining how AI safety governance is implemented across jurisdictions.

The following scenarios present three possible pathways for this institutional transition. Scenario A examines the continued expansion of international cooperation through the AI Safety Institute network. Scenario B considers a future in which frontier AI companies play the leading role in securing increasingly capable AI systems. Scenario C explores the growing integration of AI safety with cybersecurity, CBRN, and national security. These scenarios are not mutually exclusive. Instead, they represent different but potentially complementary directions for the future development of AI

safety governance.

**Table. Tracking Previous Scenario Predictions**

| May 2026 Prediction | Observations (May 1–Jul 8) | Assessment | Link to August Scenario |
|---|---|---|---|
| **Scenario A. Evaluation-Centered Operational Governance**: Evaluation and verification processes will become the core mechanism that makes governance operational. | AI evaluation held at 4th–5th centrality. EU AI Act high-risk obligations took effect Aug 2; agent safety benchmarks confirmed under 40% full-compliance task completion. | Realized: evaluation mechanisms are now operating concretely in both policy and industry practice. | Carried forward into Scenario B (The Frontier Race to Secure Agentic AI) through its red-teaming and benchmark elements. |
| **Scenario B. Behavioral Risk Governance for Agentic AI**: The focus of evaluation will expand from model outputs to long-term behavior, requiring continuous monitoring and intervention mechanisms. | AI agent held at 8th–9th centrality, and a widely cited report found models behave more safely during evaluation than after deployment, confirming the anticipated evaluation-deployment gap. | Partially realized: the anticipated behavior gap was confirmed, but continuous monitoring mechanisms are not yet institutionalized. | Carried forward directly into Scenario B (The Frontier Race to Secure Agentic AI), reaffirmed through the discussion of the capability–safety gap. |
| **Scenario C. Fragmented Governance and Distributed Accountability**: Cooperation will expand as responsibility is distributed among governments, AI Safety Institutes, companies, and independent evaluators. | AISI rose to 1st across all centrality measures, and bilateral MOUs expanded (Korea–Canada, Australia–Canada), though responsibility appears to concentrate around AISI rather than distribute evenly. | Partially realized: cooperation expanded as predicted, but concentrated around AISI rather than genuinely distributing across actors. | Carried forward into Scenario A (Institutionalized Multilateralism), with the focus narrowing from "distribution" to "AISI-centered cooperation." |

### 4.1 Scenario A. Institutionalized Multilateralism: The AI Safety Institute Network as the Central Hub

This scenario is based on AISI, Governance, Transparency, and MOU. It describes a future in which the International Network of AI Safety Institutes develops into a key mechanism for international AI safety governance. Rather than relying on a single international agreement, cooperation continues to expand through bilateral and multilateral memoranda of understanding, such as the agreements between Korea and Canada and between Australia and Canada. These agreements support cooperation in AI evaluation methodologies, incident reporting, and personnel exchanges. As cooperation expands, national AI Safety Institutes increasingly conduct AI evaluations jointly or recognize one another's evaluation results, improving efficiency and strengthening international cooperation.

The main strength of this scenario is its flexibility. The network can expand gradually without requiring agreement from all countries, allowing cooperation to develop as additional participants join. This approach also supports the gradual development of common evaluation practices and mutual confidence among participating institutions.

The main limitation is that the network depends on voluntary participation. Changes in international relations or the withdrawal of key participants could reduce the effectiveness of cooperation. In addition, differences in transparency across participating institutions may remain. While cooperation among AI Safety Institutes may continue to strengthen,

variations in the public disclosure of evaluation methods and results could affect public trust and the consistency of governance across countries.

### 4.2 Scenario B. The Frontier Race to Secure Agentic AI

This scenario is based on Anthropic, OpenAI, AI agent, Benchmark, and Mythos. It describes a future in which frontier AI companies increasingly compete to improve the safety of agentic AI systems alongside advances in model capabilities. As AI agents are deployed to perform more complex and autonomous tasks, companies continue to expand internal red teaming, adversarial benchmarking, and phased model releases. Recent examples include Anthropic's Mythos and Fable model families, which introduced different deployment tiers for general and higher-risk use cases. At the same time, independent evaluation benchmarks continue to expand alongside company-led evaluations, strengthening the overall ecosystem for assessing the safety and reliability of AI agents after deployment.

The main strength of this scenario is its speed. Frontier AI companies can improve safety measures more rapidly than regulatory processes, while competitive pressure encourages continuous investment in AI evaluation and system safety. This can help reduce the gap between performance observed during evaluations and behavior after deployment.

The main limitation is that AI safety capabilities may become increasingly concentrated within a small number of frontier AI companies. As these companies establish their own evaluation methods and safety practices, smaller developers may face greater challenges in adopting comparable standards, while independent verification of company-led evaluations may remain limited. In addition, the Mythos case demonstrates that AI safety decisions are increasingly influenced by national security policies, including export controls. As a result, frontier AI companies will need to address both technical and geopolitical considerations when developing and deploying advanced AI systems.

### 4.3 Scenario C. Securitized Safety: Cyber, CBRN, and Autonomous Weapons Converge

This scenario is based on Cybersecurity, National security, CBRN, Autonomous weapons, and Deepfake. It describes a future in which AI safety becomes more closely integrated with national security policy. Governments increasingly introduce pre-release reviews of advanced AI models for cybersecurity and national security purposes, expand export control measures, and treat evaluation results related to CBRN capabilities, autonomous weapons, and large-scale deepfake generation as sensitive information.

Developments during the reporting period illustrate this trend. Following the release of Moonshot AI's open-weight Kimi K3 model in July 2026, which the company stated performed competitively with Anthropic's Fable 5, the White House accused Moonshot AI of conducting an internal model distillation effort using export-controlled NVIDIA chips obtained through Thailand. The U.S. Treasury Department also indicated that sanctions and designation on the Entity List were under consideration. While the Mythos case demonstrated how export controls could restrict the deployment of advanced AI models by U.S. companies, the Kimi K3 case illustrates how similar measures may also be applied to

limit foreign access to advanced computing resources.

International discussions have also reflected this growing emphasis on security. During the reporting period, the United Nations convened its first dedicated global dialogue on AI governance, bringing together discussions on deepfake harms, lethal autonomous weapons, and broader AI safety issues. Although these efforts represent an important step toward international cooperation, achieving consensus on security-related AI governance remains a significant challenge. Overall, this scenario suggests that AI safety governance is becoming increasingly influenced by national security considerations, particularly in areas involving cybersecurity, CBRN, and military applications of AI.

## 5. Strategic Directions

The findings of this report suggest that the international AI Safety Institute network is becoming an increasingly important element of global AI safety governance. Korea AISI and other national AI Safety Institutes should strengthen existing bilateral cooperation by expanding shared technical infrastructure, including common AI evaluation methodologies, mutual recognition of evaluation results, coordinated red teaming, and joint incident reporting mechanisms. Strengthening these practical cooperation mechanisms will support more consistent and effective international AI safety governance.

As agentic AI systems become more widely deployed, differences between evaluation results and real-world system behavior are likely to become more significant. AI Safety Institutes should therefore expand post-deployment monitoring and behavior-based evaluation methods in addition to existing pre-deployment evaluations. Continued cooperation with frontier AI companies will also be important to improve the transparency and accountability of phased model releases and other AI safety practices.

The growing convergence of Cybersecurity, CBRN, and national security also highlights the need for closer international cooperation in handling sensitive AI safety information. AI Safety Institutes should develop secure mechanisms for sharing evaluation results and technical information while respecting national security requirements. At the same time, international cooperation should seek to ensure that export control measures do not unnecessarily restrict collaboration on AI safety evaluation and risk assessment.

Taken together, these priorities indicate that continued investment in common evaluation standards, post-deployment monitoring, and secure international cooperation will be essential for strengthening AI safety governance. Building these operational capabilities will support more reliable AI evaluations and more effective international coordination as AI systems continue to advance.

## Appendix A. Temporal Changes in Network Centrality Rankings

### 1. Weighted-degree centrality

| Rank | 2025 Aug | 2025 Nov | 2026 Feb | 2026 May | 2026 Aug | Change (May to August) |
|---|---|---|---|---|---|---|
| 1 | risk | us | regulation | governance | aisi | ▲7 |
| 2 | uk | risk | aisi | evaluation | anthropic | ▲3 |
| 3 | aisi | anthropic | evaluation | us | cybersecurity | ▲14 |
| 4 | act | regulation | safety | safety | aievaluation | ▼2 |
| 5 | llm | safety | risk | anthropic | openai | ▲11 |
| 6 | research | gpt | security | security | aisafety | ▼2 |
| 7 | eu | aisi | framework | risk | governance | ▼6 |
| 8 | us | china | deepfake | aisi | aiagent | ▲6 |
| 9 | evaluation | meta | content | regulation | transparency | ▲10 |
| 10 | anthropic | claude | control | china | mythos | New |
| 11 | china | test | child | policy | gpt | New |
| 12 | test | uk | nist | deepfake | benchmark | ▲8 |
| 13 | standard | framework | ecosystem | eu | alignment | New |
| 14 | security | research | act | ai agent | mou | New |
| 15 | report | agent ai | report | uk | nationalsecurity | New |
| 16 | transparency | eu | measurement | openai | aimodel | New |
| 17 | system | openai | agent ai | cybersecurity | frontierai | New |
| 18 | singapore | evaluation | experiment | agent | CBRN | New |
| 19 | assessment | report | us | transparency | llm | New |
| 20 | misuse | security | standard | benchmark | uk | ▼5 |

### 2. Eigenvector centrality

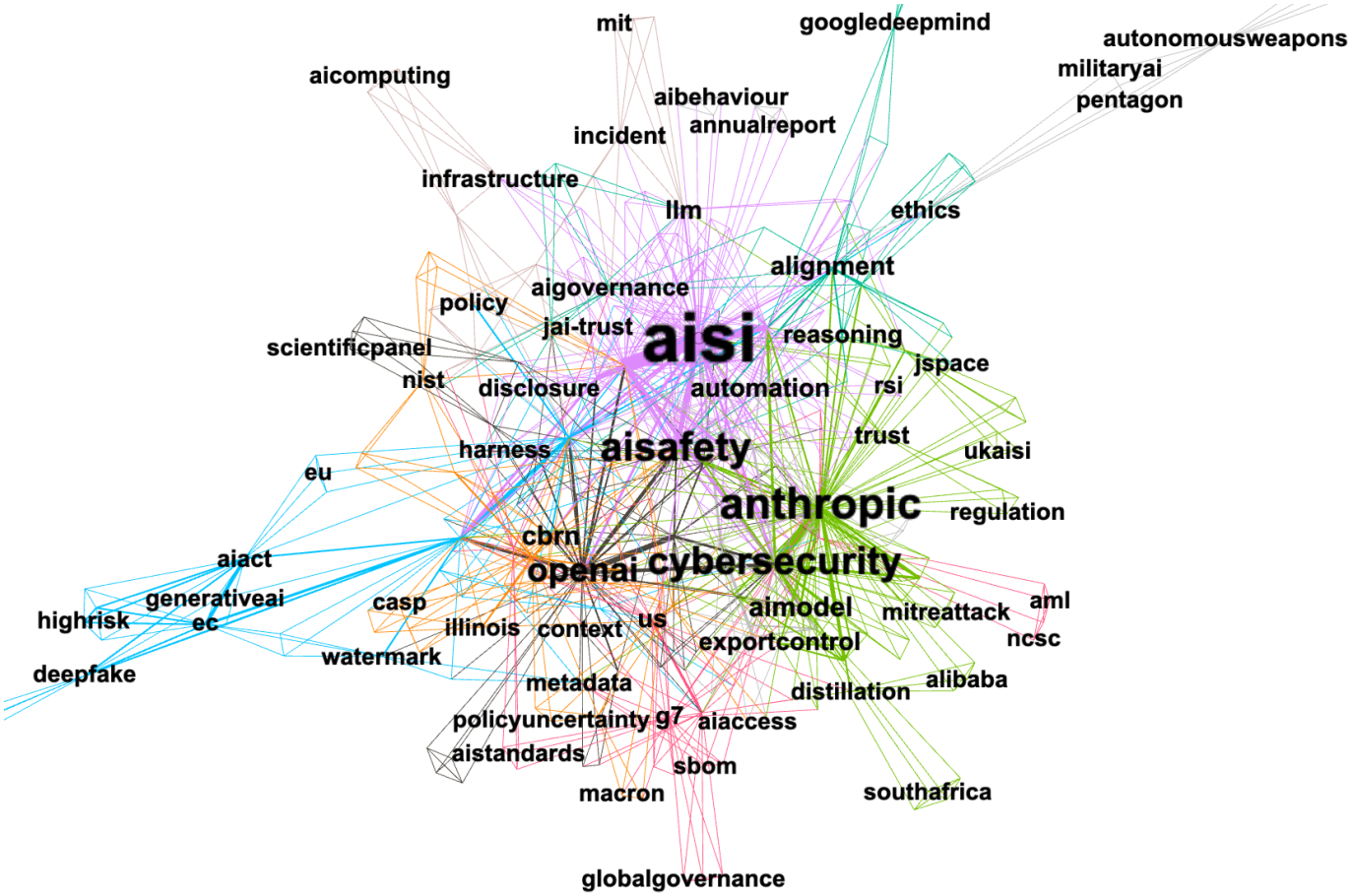
| Rank | 2025 Aug | 2025 Nov | 2026 Feb | 2026 May | 2026 Aug | Change (May to August) |
|---|---|---|---|---|---|---|
| 1 | risk | framework | regulation | governance | aisi | ▲8 |
| 2 | aisi | test | risk | evaluation | anthropic | ▲4 |
| 3 | uk | risk | security | safety | aisafety | 0 |
| 4 | act | us | evaluation | risk | cybersecurity | ▲8 |
| 5 | research | safety | framework | us | aievaluation | ▼3 |
| 6 | test | assessment | aisi | anthropic | openai | ▲8 |
| 7 | eu | innovation | safety | security | governance | ▼6 |
| 8 | evaluation | regulation | report | regulation | aiagent | ▲7 |

| | | | | | | |
|---|---|---|---|---|---|---|
| 9 | standard | ai system | test | aisi | mou | New |
| 10 | llm | terror | experiment | policy | transparency | ▲3 |
| 11 | security | security | deepfake | china | mythos | New |
| 12 | us | evaluation | child | cybersecurity | aimodel | New |
| 13 | china | exploitation | control | transparency | redteaming | New |
| 14 | anthropic | safeguard | system | openai | CBRN | New |
| 15 | framework | standard | trend | ai agent | nationalsecurity | New |
| 16 | assessment | research | governance | deepfake | uk | ▲3 |
| 17 | openai | weapon | action plan | agent | gpt | New |
| 18 | system | project | cooperation | standard | benchmark | ▲2 |
| 19 | frontier ai | report | ai system | uk | resilience | New |
| 20 | transparency | cyber security | agent ai | benchmark | internationalcooperation | New |

### 3. Betweenness centrality

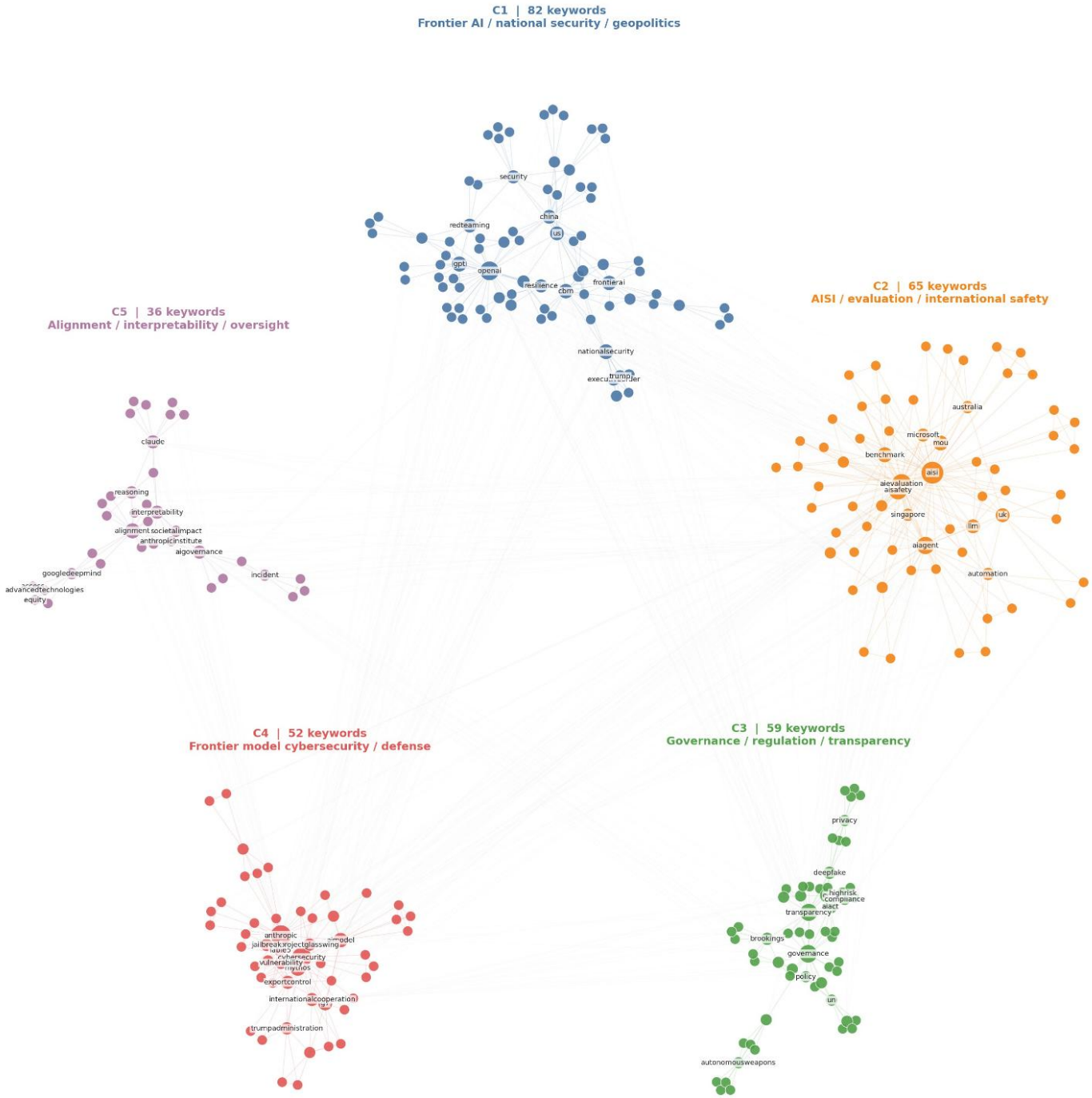
| Rank | **2025 Aug** | **2025 Nov** | **2026 Feb** | **2026 May** | **2026 Aug** | **Change** (May to August) |
|---|---|---|---|---|---|---|
| 1 | risk | us | aisi | governance | aisi | ▲7 |
| 2 | uk | regulation | us | evaluation | anthropic | ▲5 |
| 3 | act | framework | uk | safety | governance | ▼2 |
| 4 | aisi | risk | safety | us | transparency | New |
| 5 | llm | safety | evaluation | security | openai | ▲7 |
| 6 | research | chatbot | eu | risk | cybersecurity | ▲9 |
| 7 | anthropic | meta | regulation | anthropic | aisafety | ▲11 |
| 8 | us | gpt | security | aisi | aievaluation | ▼6 |
| 9 | china | innovation | openai | regulation | aiagent | ▲1 |
| 10 | security | claude | risk | ai agent | deepfake | ▲3 |
| 11 | report | deregulation | content | uk | ethics | New |
| 12 | test | evaluation | deepfake | openai | alignment | New |
| 13 | eu | legislation | report | deepfake | aimodel | New |
| 14 | responsibility | china | gpt | china | security | ▼9 |
| 15 | gpt | report | control | cybersecurity | mythos | New |
| 16 | canada | responsible ai | act | eu | frontierai | New |
| 17 | transparency | research | australia | behavior | llm | New |
| 18 | openai | agent ai | election | ai safety | g7 | New |
| 19 | evaluation | aisi | nist | cyber | googledeepmind | New |
| 20 | assessment | benchmark | guideline | agent | autonomousweapons | New |

## Appendix B. Network Topology




*Note. This topology shows the connections among key keywords across the overall network.*

## Appendix C. Sub-Network Topology




*Note. These topologies are intended to derive sub-networks by applying density-based clustering, and to analyze trends at the sub-network level based on the resulting topologies.*