AI Safety Forecasting Methodology

1. Overview

AI Safety Forecast Report, published quarterly by Korea Institute for AI Safety (K-AISI), functions as a strategic reference point for the identification of emerging risk factors associated with the accelerating advancement of artificial intelligence technologies. It further contributes to the formulation of proactive, globally coordinated response strategies. With rapid progress in generative AI and the pursuit of artificial general intelligence (AGI), growing concerns have emerged regarding malfunctions, malicious use, and systemic risks. These dynamics generate considerable uncertainty and introduce complex risks across societal, economic, and national security domains.

Against this backdrop, the report is designed to deliver regular and structured foresight into the evolving AI risk landscape. It systematically analyzes global trajectories in AI safety-related policy and governance, projects plausible future scenarios, and strengthens mechanisms for early detection and warning of potential threats. Through these efforts, the report supports the development of a robust anticipatory response framework and provides actionable strategic insights to enable effective collaboration among governments, industry stakeholders, and civil society in shaping a resilient and trustworthy AI safety ecosystem.

2. Forecasting Methodology

Forecasting in the field of AI safety leverages both quantitative and qualitative methodologies to anticipate future risks and developments. Quantitative approaches involve the systematic collection and analysis of data to extract trends, identify leading indicators, and determine the key driving factors influencing AI safety trajectories. In parallel, qualitative methods, such as expert elicitation and structured group deliberation, are employed to explore complex, uncertain, or value-laden issues that are not easily captured through empirical data alone. By integrating these methods, the forecasting process enables the formulation of plausible and policy-relevant future scenarios, which serve to inform early warning systems and strategic preparedness in the face of evolving AI technologies.

2.1 Weekly Issue Selection

As the first step in AI safety forecasting, a set of 8 to 10 key issues is selected on a weekly basis. This selection is not conducted randomly; rather, it is carried out through a structured process involving specialized research firms responsible for trend monitoring, as well as PhD-level experts in AI safety policy. To ensure a balanced perspective and avoid overemphasis on any particular domain, the selected issues are categorized as evenly as possible across four areas:

policy, industry, technology, and others.

2.2 Keyword Extraction from Selected Issues

For each selected issue, five representative keywords are extracted. This process is conducted by a team of PhD-level experts who engage in brainstorming and other collaborative techniques to identify keywords that best capture the core aspects of each issue. To ensure consistency in the dataset, a data refinement process is carried out in parallel—for example, converting verbs such as “evaluate” into their nominal forms like “evaluation.” This step serves to establish a standardized foundation for subsequent data analysis.

2.3 Data Preprocessing for Network Analysis

To facilitate network-based analysis, the extracted keywords associated with each issue undergo a structured data preprocessing phase. In this process, each set of five keywords is transformed into an edge list format, enabling the construction of a fully connected undirected network with a graph density of 1. This implies that every keyword within a given issue is pairwise linked to all others, resulting in a total of ten unique edges per issue. Subsequently, all keyword networks extracted over a specified time period are integrated into a single large-scale network, forming the foundational structure for comprehensive analysis.

2.4 Centrality Analysis for Identifying Key Driving Factors

Following data preprocessing, centrality analysis is performed on the keyword network to identify key driving factors that influence AI safety trends at the micro, meso, and macro levels. The analysis focuses on three centrality measures: weighted-degree centrality, eigenvector centrality, and betweenness centrality.

Weighted-degree centrality is used to assess the influence of individual keywords at the micro level by quantifying the total strength of their direct connections. Eigenvector centrality captures the meso-level importance of a keyword by evaluating its connectivity to other influential nodes within the network. Betweenness centrality, on the other hand, reflects the macro-level role of a keyword in bridging different parts of the network, indicating how central it is in facilitating information flow across the entire structure.

The rationale for applying network analysis in the forecasting process lies in a key principle of futures studies: while individual actors or elements may change rapidly, the network or system as a whole tends to exhibit greater structural stability. In this sense, key driving factors identified from a holistic, system-level perspective are more likely to exert sustained influence in the near future.

2.5 Expert-Based Scenario Forecasting Grounded in Key Driving Factors

A team of PhD-level experts specializing in AI safety policy conducts scenario forecasting to anticipate plausible

developments in the near future. Central to this process is the use of previously identified key driving factors as the foundation for constructing logically coherent and contextually grounded future scenarios. Through structured brainstorming and other collaborative foresight techniques, the expert group explores a range of potential futures, evaluates their plausibility, and selects the most likely scenario to put forward.

In addition to scenarios driven directly by the identified key factors, the experts also assess secondary yet meaningful keywords that were not classified as primary driving factors but may exert significant influence. This approach enables the anticipation of emerging issues that might otherwise be overlooked, thereby enhancing the robustness and forward reach of the overall forecasting framework.

3. Limitations

While this approach aims to enhance the reliability and validity of forecasting by integrating quantitative data with expert-driven qualitative insights, several limitations remain. First, the input data used for forecasting may not fully reflect the rapid pace of change in technology, regulation, and the geopolitical environment, and may be constrained in terms of timeliness, accuracy, and availability. Second, although expert judgment is valuable for interpreting complex uncertainties, it is also susceptible to cognitive biases and groupthink. Third, the scope of trend analysis is limited to overseas developments, explicitly prioritizing the activities of global AI Safety Institutes while excluding domestic (Korean) trends.

4. Significance

Despite these limitations, the AI Safety Forecasting methodology offers several meaningful contributions. First, by integrating trend analysis based on quantitative data with qualitative expert judgment, it enables the multi-layered identification of complex risk factors that are difficult to capture through a single method. Second, the application of network centrality analysis allows for the systematic extraction of key driving factors across technology, policy, and industry domains, thereby supporting evidence-based scenario design. Third, by incorporating expert-based scenario forecasting, the methodology facilitates exploratory approaches to identifying potential risks and emerging issues that go beyond conventional data-driven analysis. This combination of structural and foresight-based insights provides a strategic foundation for enabling collaborative and proactive responses among governments, industries, and civil society actors in a rapidly evolving AI landscape.