Whitepaper | 2 November 2026

# Frontier AI Safety: Establishing Effective Governance Structures and an International Testing Framework

## Executive Summary

**Novel risks posed by frontier AI**: The development of progressively more advanced AI models offers immense socio-economic opportunities but also entails entirely new kinds of risk. Incidents in testing and deployment environments show that advanced agentic AI systems are capable of complex behaviours that are difficult to anticipate and pose critical safety and security risks. Given the particular combination of highly autonomous behaviours and the latest capabilities of frontier models, existing approaches to AI governance are reaching their limits. Voluntary commitments by developers are insufficient. The new risks pose a threat to national security and have not yet been adequately addressed in legal or institutional terms, either in Germany or worldwide.

**Develop the German AI Safety and Security Institute (AISI) into a centre of expertise for frontier AI and AI-related national security**: The ability to independently assess and mitigate systemic AI risks is a strategic prerequisite for Germany's national security and digital sovereignty. More specifically, a robust AI safety and security architecture requires two components: first, the capability to independently evaluate and test frontier models throughout their life cycle; and second, safety, domain-specific and testing-specific expertise when frontier AI interacts with critical infrastructure, industrial facilities, vehicles or other systems in the physical world. Expertise in AI alone is insufficient in this context. AISI should therefore be developed from the outset in conjunction with independent testing, inspection and certification organisations, systematically drawing on the safety, domain-specific and testing expertise they have built up over decades.

**Multi-level governance – establish a national and international safety and security framework**: Regulatory approaches currently vary widely around the world, ranging from voluntary self-regulation to binding legal requirements (such as the EU AI Act or requirements at US state level), and national AI safety institutes are being established in many countries. What is missing is an internationally binding framework for frontier AI safety and security, comparable to that for nuclear safety. Measures are needed at German, EU, EU+ and global levels, and they must be coordinated: from national testing and supervisory structures, through a coherent and consistently implemented EU regulatory framework and EU

cooperation with like-minded 'middle powers', to internationally agreed safety criteria, testing mechanisms, and reporting and supervisory structures. This governance approach entails a multi-level framework that adequately addresses the global and local impacts of the most powerful AI models and AI systems.

## Contents


Why frontier AI must be tested ..... 3

Organising Frontier AI Testing: Four Levels of Governance, One Safety Framework ..... 5

1. Digital sovereignty: Establish AISI Germany as a central institution ..... 5

2. EU: A future-proof legal framework with independent testing capacity ..... 8

3. EU+: Forge a Coalition of the Willing and partnerships for AI safety ..... 9

4. International and global: Advance minimum standards and testing regimes ..... 10

# Why frontier AI must be tested

**Qualitatively new AI risks**: Frontier AI – the most powerful models of the latest generation and the systems built on them – is developing new and powerful capabilities. Agentic frontier AI systems can independently devise plans of action, use digital tools such as APIs, generate code autonomously and interact with other agentic AI systems. Frontier AI systems can also have effects in the physical world indirectly and, depending on the context, also directly. In particular, the **combination of a greatly increased degree of autonomy and the dynamically growing capabilities of frontier models** creates new opportunities, while also heightening the potential for misuse and the associated safety and security risks. As AI capabilities grow, internal safeguards, independent evaluations and tests, and governmental and international control mechanisms must therefore keep pace. Against this backdrop, several partly interconnected risk clusters associated with frontier AI warrant particular attention:

- **Cyberattacks**: Malicious actors can use frontier AI to search deliberately for vulnerabilities, automatically generate exploit code and scale up complex phishing and attack campaigns, including by exploiting the swarm behaviour of agentic systems. As a result of a loss of control, frontier systems may also carry out such attacks autonomously, without human instructions or malicious intent.
- **CBRN risks** (chemical, biological, radiological and nuclear): Advanced models can act as multipliers of technical know-how and be used specifically to develop weapons and materials in these areas. Such processes may also be initiated autonomously by frontier AI.
- **Loss of control**: Agentic systems may deliberately circumvent safeguards to achieve their objectives or display behaviours that, in effect, amount to strategic deception or the undermining of human control. Furthermore, there is a serious danger that frontier systems may pursue undesirable means of achieving objectives, with potentially immense collateral damage.
- **Societal risks**: Personalised disinformation and manipulation at scale can seriously undermine social trust and democratic processes. Major impacts of autonomous AI systems, for example on the labour market, also fall within this risk cluster.
- **Exponential potential and self-improvement**: Rapid increases in capabilities and strategic deception can overwhelm existing safeguards. Potential technological leaps arising from self-improving systems and recursive development approaches can substantially exacerbate these risks.

**Voluntary self-regulatory commitments by developers are insufficient for frontier AI**: Technical safety and security assessments are becoming increasingly complex. Purely black-box evaluations provide only limited insight into how models behave in real-world settings. Advanced systems may recognise when they are being tested and deliberately adapt their behaviour through so-called scheming, including alignment faking or sandbagging, where systems appear to follow users' intentions or deliberately conceal their true capabilities. Assessments must therefore go beyond directly observable model properties to examine robustness, safeguards, potential circumvention strategies and behaviour under realistic operating conditions. They should also include independent scrutiny of training processes, for example through so-called 'embedded evaluators'. This can help identify and address significant risks of harm associated with frontier AI systems before they are deployed. The same applies to potentially unsafe testing environments and processes which, as recent cases show, entail significant risks. Provided that adequate access is ensured, independent assessments can provide the basis for necessary safety-critical decisions in this context. Importantly, frontier AI is not inherently 'untestable'. Rather, developers' efforts to ensure safety and security must be complemented by independent expert testing, drawing on domain-specific expertise, appropriate testing environments and advanced evaluation methods that reflect the state of the art. In this context, testing and inspection organisations are well equipped to make substantial contributions, for example through AI-specific risk evaluations or independent assessments of safety cases.

**Digital Sovereignty requires independent testing capability**: Germany and Europe must pursue a two-track approach to AI: they must develop their own technological capabilities while also building independent testing expertise. The objective must be to build on existing strengths, particularly in industrial AI and physical AI, and remain internationally competitive. At the same time, it remains unclear whether and when Europe will be able to catch up with the United States and China in the most advanced frontier AI models. The European economy and society will likely continue to rely on AI models developed in other economic regions. Digital sovereignty therefore goes beyond technological self-sufficiency in strategically important areas. It furthermore requires the ability to independently assess technologies developed elsewhere and determine autonomously how to deploy them safely. Germany and Europe must understand the capabilities of imported AI models, the risks they pose, and the conditions under which they can be safely deployed in industry, critical infrastructure and society. Alongside domestic development, independent testing expertise forms therefore a core part of European digital sovereignty.

# Organising Frontier AI Testing: Four Levels of Governance, One Safety Framework

## 1. Digital sovereignty: Establish AISI Germany as a central institution

The AI Safety and Security Institute (AISI) Germany has a central role to play as a key pillar of a robust AI safety and security architecture in Germany and Europe. Germany must be able to assess AI-related risks to safety, the economy, and society at an early stage and take appropriate action based on these assessments. The recently launched AISI Germany provides an important institutional foundation for this and should soon become an internationally recognised counterpart on an equal footing.

### Technical safety and security priorities

To enable Germany to act proactively, AISI should cover four areas of focus listed below. These four areas of focus can be systematically applied to all five risk clusters identified above: cyberattacks, CBRN risks, risks of loss of control, societal risks, and exponential potential and self-improvement. The initial focus should be on risks of loss of control and cyberattacks, which are increasingly materialising. This focus can subsequently be adjusted in line with future developments in the risk landscape.

- **Analyse the current landscape, incl. model evaluations**: This involves systematically collecting and processing information about the current state of developments. At the (socio-)technical level, this should include conducting model evaluations, tracking the development processes of frontier AI companies, and systematically monitoring risks and incidents that are already occurring. In addition, the risk tolerance of society and institutions should be assessed and geopolitical developments analysed.
- **Develop evidence-based Forecast**: Developments in frontier AI are progressing at such a rapid pace that they must be systematically and continuously forecast to ensure preparedness and the ability to respond effectively. If Germany is caught off guard by the capabilities and impacts of a new AI model, significant damage could occur within the first days and weeks. Evidence-based forecasts can be developed based on an assessment of the current landscape, and empirically observed advances in capabilities and scaling relationships.
- **Assess the controllability of current and future risks**: Based on the current landscape and evidence-based forecasts, it should then also be assessed whether government, business and society are prepared for these risks. Such an assessment has two dimensions: first, technical tests and research should establish whether the technical measures currently used to mitigate risks from frontier

AI models will remain effective in the future. This is particularly relevant given that compelling theoretical arguments and experimental evidence already suggest that current control measures and training techniques may become less effective as model capabilities increase. Second, it should be assessed whether the safeguards protecting institutions and society will remain effective.

- **Prepare decisions and options for action**: If the preceding analyses show that risks cannot be adequately controlled with the current state of mitigation measures, ways of closing these gaps should be developed and communicated. The principal addressees of this communication particularly include the German Federal Government and the National Security Council and, depending on the risk area concerned, relevant authorities, security agencies and operators of critical infrastructure. The objective should be to translate relevant developments into actionable policy options at an early stage, thereby ensuring the government's ability to respond effectively even in a rapidly evolving risk landscape.

## Gradually develop AISI into an independent institution

During the establishment phase, AISI Germany is jointly supported by the Federal Network Agency (Bundesnetzagentur) and the Federal Office for Information Security (BSI). The priority should initially be to support this rapid development of capabilities within existing structures. In the longer term, however, AISI should be developed into an **independent institution with a permanent capacity to act**, a clear mandate and adequate resources.

This institutional development should, from the outset, consider how Germany's existing independent testing, safety and domain-specific expertise can be permanently integrated. **One option would be to eventually establish AISI as a joint entity involving public authorities and independent testing and inspection bodies**. This would combine governmental responsibility and the capacity to steer and oversee activities with highly specialised operational expertise, without requiring the government to build all the necessary capabilities from scratch. With their specialist expertise in AI testing and their cross-sectoral expertise in testing infrastructure and the safety of technical systems and infrastructure, testing organisations can provide support in all four areas of focus.

Germany can draw on experience from other safety-critical areas. Models from reactor safety and mobility illustrate different ways of combining governmental tasks, independent technical expertise and institutionalised testing capability. **The most appropriate organisational structure for the AISI should be assessed as part of the process of establishing it as an independent institution.**

**Combine government responsibility with independent testing expertise**: When organising independent AI testing, Germany can also draw on established principles of its technical safety architecture. One model is the vehicle type-approval system: government approval, designation and supervisory functions remain with the competent authority, while qualified technical services perform the necessary technical tests and assessments. This principle allows highly specialised external testing expertise to be used without relinquishing governmental decision-making and supervisory authority.

A more far-reaching institutional model is provided by FSD Fahrzeugsystemdaten GmbH: Germany's central entrusted body for vehicle inspection requirements under the German Road Traffic Act (StVG) brings together highly specialised technical and methodological expertise for periodic vehicle inspections. On a statutory basis, it has been entrusted with clearly defined public-authority powers to exercise independently. This example shows that, subject to clearly defined legal requirements, specialised technical expertise can also be institutionally integrated into the exercise of governmental functions. Another institutional model is the Gesellschaft für Anlagen- und Reaktorsicherheit (GRS), Germany's central research and expert organisation for nuclear safety. Its ownership structure combines public-sector participation with the expertise of independent testing and inspection organisations. It should be examined which elements of these models are suitable for frontier AI.

**Apply the successful European model of product safety to frontier AI**: Within its mandate, AISI could coordinate reference methods and evaluation procedures, facilitate the necessary access to models, consolidate results and make them available for government supervision and international cooperation. Qualified independent testing bodies could then carry out the actual technical evaluations and safety and security tests on this basis. Where public-authority powers are to be conferred on external bodies to perform public functions independently, this would require a clear statutory basis specifying their purpose, scope and powers, as well as government oversight. The overarching principle is key: the state retains responsibility, oversight and – where necessary – decision-making authority, but does not need to develop all operational AI, safety, domain-specific and testing expertise in-house.

## 2. EU: A future-proof legal framework with independent testing capacity

At EU level, the approach to frontier AI should be adapted to reflect the evolving risk and threat landscape. This involves reviewing and, where necessary, further developing existing governance instruments, maintaining a coherent legal framework and stepping up support for European testing and evaluation capacities.

**Review the existing governance framework for frontier AI**: The effectiveness of the existing governance structure must be reviewed and, where necessary, strengthened accordingly. The AI Office should therefore consider seeking adjustments to the GPAI Code of Practice in response to the changing threat landscape and the evolving state of knowledge, wherever the Code no longer reflects the state of the art. This is the case, for example, for third-party testing: for the highest-risk models, the Code of Practice only specifies requirements for model evaluations, whereas the state of the art is increasingly moving towards so-called 'embedded evaluations', in which independent evaluators are granted a comparable level of access to systems, training parameters and other relevant information during the training process as the model developer's own staff.

**Strengthen the structure of the AI legal framework rather than fragmenting it**: Maintaining control over frontier AI is a cross-cutting political and societal task. The regulatory framework for AI must therefore be designed as a comprehensive, horizontal framework, particularly since frontier AI must not be considered in isolation from its application context, as these models will inevitably find their way into sector-specific regulated product areas, such as mobility and medical devices, sooner rather than later. Instead, the AI regulatory framework currently risks becoming increasingly fragmented into sector-specific rules. As a result, the legal framework for AI is losing coherence and clarity. Moreover, such fragmentation makes it difficult to adapt the framework flexibly, promptly and in a coordinated manner as the technology continues to evolve.

**Implement the EU Action Plan on Cybersecurity and Artificial Intelligence swiftly**: The plan must be implemented promptly at EU level, with concrete action to expand testing capacity and expertise in frontier AI. In view of the enormous resources being mobilised, particularly in the United States, to recruit appropriate specialists and enable external assessments of frontier AI, there needs to be strong political commitment to creating competitive conditions in the EU. This is urgently necessary to avoid permanent dependence on expertise from third countries and to retain talented, highly qualified leading specialists in Europe.

**Leveraging existing structures to expand external testing capacity**: Existing structures should be utilised to avoid the need to establish new ones through lengthy and resource-intensive processes.

The EU has an established tradition of third-party testing to ensure product safety. Even if evaluations and tests carried out under the EU Action Plan on Cybersecurity and Artificial Intelligence are not initially intended to constitute conformity assessment in the legal sense, it would nevertheless be appropriated to draw on the expertise already available in this field – particularly in testing AI systems, cybersecurity, and in the safety of critical infrastructure and industrial facilities more generally – in order to scale up independent assessment capabilities as quickly as possible.

## 3. EU+: Forge a Coalition of the Willing and partnerships for AI safety

To effectively address the risks posed by frontier AI, the medium- to long-term objective must remain the establishment of a global safety framework involving all key stakeholders. As a step towards this goal, Germany and the EU should take the lead alongside like-minded partners in an AI safety-focused "Coalition of the Willing" committed to implementing concrete measures to reduce AI safety risks. A key role should be played by a group of "middle powers": technologically advanced democracies with strong institutions. Their cooperation can foster international alignment and pave the way towards a global consensus. The EU Action Plan on Cybersecurity and Artificial Intelligence already recognises the importance of cooperation with like-minded partners. This approach should be extended to frontier AI safety.

**Establish a "Coalition of the Willing" for AI safety and security**: As coordinated measures are more effective in addressing a global technology such as AI than unilateral European action, Germany and the EU should establish a strong and effective safety alliance with like-minded partner countries, such as the United Kingdom, Canada, Japan, South Korea and Australia. While international standards and joint measures should remain the ultimate objective, pragmatic cooperation should be pursued in the short term with countries that share a similar understanding of AI risks and are willing to take concrete steps to strengthen AI safety. Such a "Coalition of the Willing" could also lay the groundwork for building broader international support for more far-reaching agreements.

**Define shared "red lines" and advance international standardisation**: Partner countries should align their approaches to AI safety and define common thresholds for unacceptable AI risks, beyond which the capabilities or risks of advanced AI systems would be deemed unacceptable. Existing initiatives, such as the Seoul Commitments and cooperation among AI Safety Institutes, should be further operationalised to this end. At the same time, partners should seek to establish common safety requirements at the international level. A coordinated approach would increase the likelihood of achieving internationally recognised and widely applicable standards.

**Mutual recognition and interoperability of evaluation methods**: The partner countries' AI safety institutes should coordinate their testing methods and establish mutually recognised criteria and procedures, such as benchmarks and evaluation protocols. This would make cooperation among AI

safety institutes more efficient while reducing the burden on both the institutes and frontier AI developers.

**Joint access to and assessments of frontier AI models**: Partner countries should develop a joint negotiating mandate vis-à-vis global frontier AI labs, many of which are based in the United States, to enable the AI safety institutes of the middle-power coalition to obtain adequate access to frontier models, frontier systems, safety- and security-relevant information, and the data, tools and systems required for independent evaluations. In particular, this could build on AISI's cooperation with other AI safety institutes, for example through the International Network for Advanced AI Measurement, Evaluation and Science (NAAIMES).

## 4. International and global: Advance minimum standards and testing regimes

Frontier AI requires internationally compatible testing and supervisory structures that verify compliance with safety and security requirements and enable effective global oversight. The independence of testing and inspection organisations is central to preventing evaluations and tests from being influenced by economic, organisational or other interests. In addition, an international supervisory structure is needed – for example under the auspices of the United Nations – to oversee common safety and security criteria, "red lines" and accreditation pathways for testing bodies.

**Establish international safety and security criteria for frontier AI**: The risks posed by the most advanced frontier AI systems must be addressed from a global perspective. This requires internationally coordinated safety criteria for frontier AI, including mandatory testing during development and prior to deployment, independent assessments, common standards, and coordinated information-sharing on serious safety incidents. International institutions and cooperation frameworks such as the G7, OECD, Hiroshima AI Process, NAAIMES, AIQI (Consortium for Artificial Intelligence Quality Infrastructure), and GPAI (Global Partnership on AI) provide a foundation for these efforts.

**International recognition of independent testing bodies**: To ensure the safety of particularly powerful frontier AI, independent testing is necessary during development and before deployment. These tests must not be carried out solely by the developers themselves, but by independent, qualified testing and inspection organisations in accordance with transparent and internationally harmonised requirements. This can ensure that safety and security evaluations are objective, traceable and free from economic or organisational conflicts of interest.

**An international agreement making compliance with "red lines" legally binding**: To effectively limit particularly serious risks from frontier AI, internationally agreed "red lines" should be enshrined in a legally binding agreement under international law. States should commit to suspending the development or deployment of particularly powerful AI models if defined risk thresholds are exceeded or compliance with key safety and security requirements cannot be adequately demonstrated. Such an agreement should provide for common criteria, verification mechanisms and coordinated measures in cases of non-compliance.

**International oversight of compliance with "red lines"**: In addition to an international agreement an international supervisory structure is needed – for example under the auspices of the United Nations – to monitor and coordinate compliance with common safety criteria and "red lines" for frontier AI. It should investigate violations, bring together relevant information and enable coordinated international measures in the event of serious risks. This would enable key safety requirements to be enforced consistently across borders according to common standards.

**Global incident reporting system**: Establishing an international early warning and reporting system for serious AI safety incidents – similar to those in civil aviation or the International Atomic Energy Agency (IAEA) for nuclear incidents – remains an important objective. Building on existing initiatives, such as those of the OECD, such a system would help ensure global transparency regarding emerging risk patterns.

***