GPAI SAFE Project Report

Technical Trustworthiness and Data Governance Assurance of Models

June 2025

This report was produced as part of the SAFE Project and presents the outcomes of activities carried out under the project.

GPAI | THE GLOBAL PARTNERSHIP ON ARTIFICIAL INTELLIGENCE

This report was planned prior to the integration of the Global Partnership on Artificial Intelligence (GPAI) and the Organisation of Economic Co-operation and Development (OECD) in mid-2024. Consequently, the report was not subject to approval by GPAI and OECD members and should not be considered to reflect their positions.

Acknowledgements

This report Technical Trustworthiness and Data Governance Assurance of Models was developed in the context of the 'Safety and Assurance of Generative AI (SAFE) Project', with the steering of the Project Co-Leads, Inma Martinez and Cyrus Hodes, and the guidance of the Project Advisory Group. The GPAI SAFE project agreed to declassify this report and make it publicly available.

Lead: Inma Martinez*

The report was written by: Prof. Ulises Cortés*, Prof. Kyoko Yoshinaga*, Merve Hickok*, Dr. Alistair Knott*, Alexandre Reeberg de Mello*, and Inma Martinez*

GPAI would like to acknowledge the tireless efforts of colleagues at the GPAI Tokyo Expert Support Center.

* Expert
** Observer
† Invited Specialist
‡ Contracted Parties by the CofEs to contribute to projects

Citation

GPAI 2024. SAFE Project Report, "Assuring AI in 2025: Overview and analysis of the activities of the AISIs and the International Network partners and recommendations on future model validations on Generative AI systems from a technical and data governance perspective (Technical Trustworthiness and Data Governance Assurance of Models)", June 2025, Global Partnership on AI.

Foreword

The Safety and Assurance of Generative AI (SAFE) Project was established as a cross-Working Group initiative based on the GPAI Work Plan 2024. The project divided its activities into two areas of research: "Technical Trustworthiness and Data Governance Assurance of Models" and "Mapping AGI Safety Solutions." The SAFE Workforce was set to be transversal, liaising with all Working Groups under GPAI 1.0. It was also formulated to serve as a coordination body to liaise with other global initiatives on responsible Generative AI set out by governments, international organizations, academia and research centres, civil society, as well as the private sector from startups to corporations.

The approaches adopted to analyse and evaluate the safety and trustworthiness of current AI Generative and Agentic AI model training are driven by the G7 Hiroshima AI Process's request to GPAI Experts of "conducting practical projects" that can help assure the safety of commercializing generative AI and supporting safe and trustworthy innovation beyond regulatory frameworks that can advance both the need of industries to remain competitive, forge a safe and trustworthy pathway to AI innovation and support government efforts towards industry standards that are based on service-led principles.

This report presents the findings related to the technical model validations performed by AI developers, AI safety agencies and other observatory labs whose mission is to verify the validity of practices, the alignment with values, principles and objectives of all science and technological innovation, and the industry requirements across all sectors to adopt innovative technologies that perform according to standards.

It presents analysis and recommendations on global efforts initiated in 2024 which were presented to GPAI members during the Paris Action Summit of February 2024, a set of ten recommendations that today, some half year thereafter, we see materialising as increased risks and challenges encountered during the training lifecycle of Generative AI models. The intricate interplay between technical complexities, ethical and legal imperatives, and the evolving regulatory landscape demands the publication of this report. It stands as an advice to how the challenges and assumptions from 2024 have shifted in response to current developments in model creation and validation. Recent legislative frameworks highlight this transformation; for example, the European Union's Artificial Intelligence Act states, "High-risk AI-based systems shall be designed and developed in such a way, including with appropriate human oversight, that they can be effectively overseen and that risks to health, safety and fundamental rights are minimized to an acceptable level" (EU AI Act, Article 9). This requires the continuation of this project beyond the findings presented within it. As evidence today has demonstrated that ensuring safety and trustworthiness is not merely a supplementary consideration but an intrinsic requirement that must be embedded throughout the entire model training and development process, necessitating a holistic and interdisciplinary approach. The corpus of GPAI Experts, through their diverse expertise, practical experience, and influential networks, are uniquely positioned to drive this effort forward for the benefit of all GPAI members, in line with the spirit and letter of these new legal frameworks, as well as industry's emerging best practices.

In the past few months, the AI research community has accounted for an accelerated progress in the capabilities and functionalities of frontier AI-based systems, which encompass Large Language Models and sophisticated diffusion models that have demonstrated their potentiality to become highly effective AI tools. However, in parallel, model training has also suffered tremendous set-backs where it comes to model alignment, model control, and deception which have turned their training into increasingly challenged environments to assure safety and trustworthiness. The very act of scaling up the capabilities of these models also proportionally scales their potential for unintended harm, thereby demanding the development and implementation of a disproportionately complex and robust safety framework.

As AI-based systems grow in power and become increasingly integrated into critical societal and governmental functions, the real-world effects on individuals and communities necessitate a profoundly responsible approach to their development, which is being currently forecasted by some AI researchers to reach a point of no return around 2027-2028 if training is not modified and adjusted to model assurance at present and the intricacies of assuring AI-based systems. This can involve combining simple generative modeling paradigms with successful network architectures, training on vast datasets, and incorporating human feedback to refine preferences. In addition, there is growing desire by governments across the globe to develop sovereign AI-based systems, and this will add further pressure to the safety baselines of such national models in order to avoid walled gardens of development that do not scale internationally if they were to be commercialised beyond local markets.

We need to arrive into a commitment that positions safety and trustworthiness of AI-based systems not as optional goals, but as intrinsic requirements—challenges that demand continued, rigorous, and collective action.

It is therefore our aim to propose in an Addendum[1] to this report our recommendations for further areas of research and monitoring around safety and trustworthiness that GPAI Experts would conduct within the GPAI remit of "*areas of interest*" as a precursor of further project milestones within the GPAI-OECD framework.

[1] The Addendum to this report was added as of September 2025.

Table of Contents

Executive Summary

The GPAI Project Safety and Assurance of Generative AI (SAFE) was launched as a cross-Working Group initiative based on the GPAI Work Plan 2024, in response to growing global concerns about the risks and opportunities posed by advanced artificial intelligence systems. Framed by the objectives set forth by the G7 Hiroshima AI Process and aligned with the commitments made at the Bletchley and Seoul AI Summits, this report addresses both the technical and ethical dimensions of Generative AI, with a focus on interoperability, validation, data governance, and societal impact. The project consists of two components: "Technical Trustworthiness and Data Governance Assurance of Models" and "Mapping AGI Safety Solutions."

This report presents the findings related to the former component, and focuses on two principal themes:

(A) The mapping of interoperability across Large Language Models (LLMs), the technical validation frameworks developed to assure their reliability, and the industrial deployment of advanced AI as a proprietary and strategic asset; and

(B) The evaluation of data governance practices and the ethical, human-centric implications of downstream applications of LLMs and Generative AI systems.

These two themes are examined jointly through five key dimensions that structure the analysis throughout the report.

Interoperability and Technical Validation of Advanced AI

As Generative AI systems rapidly evolve and proliferate across commercial and public sectors, the need for harmonized validation standards and robust technical assurance has become urgent. The report analyses methodologies implemented by AI Safety Institutes (AISIs), research organizations, and developers to evaluate the performance and resilience of LLMs. These include:

- Mechanistic Interpretability: Ensuring that internal model processes can be interpreted and validated.
- Programmatic Trustworthiness: Auditing data pipelines, training protocols, and operational environments.
- Performance Metrics: Evaluating accuracy, fluency, and relevance of model outputs.
- Failure Audits: Systematically diagnosing output anomalies and failure modes.
- Adversarial Robustness: Testing model resilience against malicious attacks and manipulation.

The report identifies industrial domains where validation is especially critical: autonomous vehicles, logistics, legal and financial services, healthcare, agriculture (robotics and drones), and aerospace. In these sectors, Generative AI and LLMs are deployed in mission-critical or sensitive contexts, requiring certification processes that meet safety, legal, and ethical standards.

Data Governance and Ethical Alignment

The SAFE Project underscores that technical robustness alone is insufficient without parallel attention to ethical and human-centric governance. A core component of the analysis focuses on data quality, lineage, integrity, and veracity—factors that significantly affect model trustworthiness. It advocates for comprehensive auditing and verification practices that go beyond traditional data protection regulations and address the full lifecycle of training and operational data.

In addition, the report highlights persistent discrepancies in the terminology and taxonomies used by technologists, regulators, and policymakers. These inconsistencies hinder policy alignment and global cooperation, especially when trying to reconcile domain-specific interpretations of concepts like "safety," "alignment," and "explainability." The harmonization of terminologies is identified as a foundational step toward effective global governance.

Mapping and Mitigating Advanced AI Risks

Recognizing the potential for Artificial General Intelligence (AGI) to pose systemic risks, the SAFE Project introduces a "defence in depth" framework. This approach outlines a suite of strategies aimed at both endogenous (self-directed) and exogenous (externally driven) risks. These include:

- Mechanistic and behavioural analysis of AI decision-making processes.
- Red teaming and simulation of adversarial attacks, including persona-modulation and jailbreaking.
- Transparency and verifiability mechanisms, including explainable AI and auditing tools.
- Containment and shutdown mechanisms ("kill switches") for high-risk systems.
- Value alignment techniques to ensure that AI objectives remain consistent with societal norms.
- These proposals are set to be presented at the Paris AI Action Summit in February 2025 and form the basis of future work in global AI safety cooperation.

Societal and Industrial Impact of Generative AI

The report draws attention to the social and economic implications of Generative AI. While governments and large enterprises have rapidly integrated these systems, small and medium-sized enterprises (SMEs) face structural and financial barriers, including high implementation costs, data privacy concerns, and lack of internal AI expertise. There is also internal resistance to adoption due to fears of job displacement and technological redundancy.

In creative sectors, Generative AI presents philosophical and practical challenges. As it increasingly encroaches upon domains traditionally reserved for human expression—writing, art, design, and music—concerns grow over originality, authorship, and the commodification of creative labor. The report advocates for clearer labelling of AI-generated content, the preservation of human-authored works, and mechanisms that allow consumers to choose between human and machine-generated outputs.

Governance Models and Global Regulatory Trends

Across jurisdictions, the report observes wide variance in AI governance models:

- European Union: Pursuing a hard law approach through the AI Act, with legally binding requirements on transparency and deepfake labelling.
- United States and United Kingdom: Emphasizing soft law and voluntary compliance, with a focus on pre-deployment model testing via AISIs.
- Japan: Utilizing soft law frameworks supported by a strong culture of compliance and sector-specific amendments to existing legislation.

These differences reflect distinct political cultures, resource levels, and regulatory philosophies. Nevertheless, the report finds common ground in the shared commitment to developing interoperable safety protocols and fostering public trust.

Conclusion

The SAFE Project concludes that the safe and ethical deployment of Generative AI requires a balanced, multi-stakeholder approach that integrates technical validation, data governance, and societal impact assessment. The development of shared standards, terminology, and collaborative frameworks is essential to preventing fragmentation and ensuring that AI technologies benefit society at large.

As Generative AI becomes embedded in critical infrastructure and daily life, governments, industry, academia, and civil society must work together to shape an ecosystem where innovation does not come at the expense of safety, accountability, and human dignity.

1. Overview of the global challenges affecting AI safety

With the rise of Generative AI and growing concerns about its impact, the Hiroshima AI Process was initiated in May 2023 to examine the opportunities and risks associated with AI, following the directives set by the Leaders at the G7 Hiroshima Summit. After ongoing discussions, including an interim ministerial meeting in September and a multi-stakeholder high-level meeting at Internet Governance Forum (IGF) Kyoto in October 2023, the "Hiroshima AI Process Comprehensive Policy Framework"—the first international framework incorporating Guiding Principles and a Code of Conduct to promote safe, secure, and trustworthy advanced AI systems—was successfully adopted at the G7 Digital & Tech Ministers' Meeting in December 2023 and endorsed by the G7 Leaders later that month.

In November 2023, AI Safety Summit was held at Bletchley Park, Buckinghamshire in UK and the Bletchley Declaration was signed by the governments in attendance, emphasizing the need for safe, trustworthy, and responsible AI development through international cooperation calls for collective efforts from governments, industry, academia, and civil society to mitigate risks such as bias, privacy concerns, and potential misuse, especially from powerful frontier AI models. Recognizing the global nature of AI challenges, the declaration advocates for risk-based policies, transparency, and sustained research to ensure AI benefits humanity while minimizing harm.

The AI Seoul Summit was held on May 21–22, 2024, advancing global cooperation on AI safety, innovation, and inclusivity achieving (1) Seoul Declaration for Safe, Innovative, and Inclusive AI which fosters international collaboration to develop interoperable AI governance frameworks, (2) Frontier AI Safety Commitments by sixteen leading tech companies, and (3) Seoul Ministerial Statement signed by 27 countries and the EU which emphasizes the importance of a global strategy focused on AI safety, innovation, and inclusivity. It highlights the need for transparency, accountability, and robust risk management for advanced AI models, while promoting AI-driven ecosystems, public sector applications, and sustainable practices.

Amid these significant developments, the AI Safety Institute (AISI) was established elsewhere, marking the start of international discussions aimed at creating a robust framework to ensure AI safety.

The global coordination of efforts to assure the safety of artificial intelligence in 2024 confirmed that the principal concerns of governments on misinformation and malicious use, bias and discrimination, privacy and surveillance, security risks, and economic and social impacts, as well as the regulatory challenges, were addressed by the different AISIs according to a variety of approaches not only driven by the difference in national resources and financial budgets, but fundamentally by the cultural mindsets of each nation, and their particular understanding of how challenges should be addressed. While still allowing for AI companies to continue the developing road maps of their Frontier and AGI models, AISIs are primordially focusing on the immediate effects of Generative AI systems derived from these models, pressured by their widespread use in society and across business sectors, who are incentivised to embed them into their internal processes.

The challenges that the AISIs face cover a wide range of issues: from how Generative AI is still perceived negatively or misunderstood by society, to the lack of adoption and acceptance by small and medium enterprises to the pressure that Generative AI exerts on human creativity, affecting the future of jobs in creative industries.

The testing of Large Language Models (LLMs) presents technical challenges that not every AISI can address, while new risk factors emerged from the public use of Generative AI and its malfunctions and infringements on data require new approaches to traditional data governance beyond copyrights and privacy, demanding data lineage and provenance, data integrity and data veracity that assure the trustworthiness of models.

1.1 Perceptions of and reactions to Generative AI in society and business

According to Ipsos, the biggest challenge for the adoption of Generative AI systems by citizens for private use is the perceived lack of benefits. Many people argue that these tools fail to offer any clear and direct usefulness in their daily lives, therefore contributing to the general perception that Generative AI systems were released to citizens with very low levels of “fit for purposefulness”, that is, solving concrete problems, and with functionalities that, though seemingly convenient, challenged the safety of using them and the trustworthiness of their outputs - generated written and visual content. While in 2023 there were bans on ChatGPT and similar AI tools in educational institutions, the trend in 2024 has been to teach students about AI literacy, critical thinking, and the ethical implications of using these tools, as well as working on developing AI policies that address issues like plagiarism, academic integrity, and data privacy.

In the EU, the AI Act includes specific requirements for AI literacy where both developers and users of Generative AI must not only understand the technology, and its risks, but also its ethical implications. Still, the International Association of Privacy Professionals (IAPP) has highlighted that there is no single definition of what constitutes AI literacy, suggesting that AI awareness programs need to address different levels of understanding for various target audiences, scenarios of use, and cultural idiosyncrasies.

The adoption and understanding of the value proposition offered by Generative AI among small and medium enterprises is still cautious, amid concerns regarding intellectual property data having to be shared with the Generative AI providers in order to develop internal bespoke solutions, and the high costs of development, still unaffordable for most SMEs. In addition, many SMEs lack the in-house expertise to understand, implement and manage AI solutions effectively. Generative AI has arrived to them not fully turned into a product, but open to be “interpreted” internally by stakeholders as a technology offering functionalities that they, in turn, have to reimagine and shape into a solution that may solve their problems or be built into a competitive attribute. At present, SMEs do not have such product development capabilities, and their IT solutions are purchased or outsourced to fully developed solutions offered to them by third parties, and this, more than anything else, is what keeps them reluctant to invest time and resources into sandboxing, testing or even piloting Generative AI solutions.

Internally, SMEs also suffer from employee resistance to Generative AI adoption, due to concerns about job displacements or the need for skills and training that they do not have access to. If senior management is unclear about developing an AI vision for their firms, or even an AI roadmap for sandboxing, it is due to the unclear understanding of the benefits that Generative AI and automation offer, which makes it difficult to reveal clear return on investment (ROI), even if potentially adoption of Generative AI could offer much higher operational efficiency.

The rapid pace of evolution that Generative AI presents makes it even harder to be adopted by SMEs, which move much slower when onboarding emerging technologies into their business processes, increasing the gap between understanding the potentiality of Generative AI and the ability to choose the most appropriate functionalities that could be turned into safe and trustworthy solutions.

1.2 Generative AI as a challenger to human creative outputs

Outside of the practicalities and efficiencies that Generative AI may be capable of offering, the fact that it is an AI system that presents formulated, and constructed outputs that are unstructured, in the form of written or visual representations, directly challenges the area of human creativity. The ease with which Generative AI seems to be able to produce these outputs devalues human skills and efforts, even when most of the outputs seem to lack ‘artistry’, resembling pieces of creative work weirdly produced by machines. This occurs because AI models, which learn from data, do not have the same lived experiences, emotions, and perspectives as humans. The result is an AI-generated art that, while technically proficient, lacks the depth, nuance, and emotional resonance of human creations. Furthermore, as AI models often draw from existing trends and patterns, this could lead to a homogenization of creative styles and a decline in truly original and groundbreaking work. Currently, the reaction of policy makers and regulators is to approach the issue via the argument of copyright and ownership of works. As AI models are trained on vast datasets, often including copyrighted material, it is unclear who owns the copyright to AI-generated art: the programmer, the user who provided the prompt, or the AI itself. This lack of clarity could create legal and ethical challenges.

Beyond the copyright issues, there is a growing concern that, as Generative AI continues to improve, the overflowing availability of AI-generated content could seriously challenge the creative industries, creating the risk of job losses if employers opt for cheaper AI-generated content over human creators, who may be forced to shift towards being "prompt engineers" or curators of AI-generated content.

The effect of Generative AI in the creative industries is bringing up a new philosophical dilemma around what creativity is and if machines should be trained to overtake this human unique capability, evolving the concerns from quality of output towards appropriateness or suitability of output. Appropriateness as in what content in society should be produced by machines, and which content should be produced by humans, and suitability as in a future scenario where machine generated content should be clearly labelled and users of AI or beneficiaries of AI should have the choice of opting for human generated content at will. On an economic level, who is to challenge that the price of human-generated content should not cost more and be more valuable and unique, when AI-generated content drives creativity towards a low-level commodity accessible to all but lacking the beauty of human craftsmanship?

Generative AI is not only dangerously at the risk of discouraging aspiring artists and undermining the perceived value of human created art, but also if AI-generated content floods the market, it could be harder for human artists to compete and earn a living, creating a decline in the diversity and originality of creative works.

1.3 Emerging foundation models that respond to AI Safety: small foundation models

While the centre stage in government discussions around AI Safety has been occupied by Large Foundation Models (LFMs) - large AI models trained on massive datasets, capable of performing a wide range of tasks like language understanding, image generation, and code creation, industry and AI researchers have been developing smaller versions of these models capable of downstreaming AI applications that are more efficient in terms of costs and energy consumption, as well as presenting lower barriers for adoption by businesses. Small Foundation Models (SFMs) are trained on less data and have fewer parameters than their larger counterparts. They can even derive downstream applications directly from on-premise customer data, that is, data sets that remain within the control and confines of their business owners, radically diminishing the need to share proprietary data with AI developers.

There are multiple methods for scale-down LLMs, and those are a viable path to increase the data and model verification, thus the system trustworthiness, by narrowing the application to a specific context. Methods like quantization, adaptation, fine-tuning, or knowledge distillation are solutions commonly adopted by developers who need to niche their application with a more robust safety mechanism. These strategies make it easier to implement safety mechanisms as the data can be curated and the model better comprehended.

The philosophy of SFMs is also adopted by neuromorphic chip manufacturers, as TrueNorth and Loihi, using new architectures (as Spiking Neural Networks) for specialized hardware. The usage of specialized hardware for specific tasks are less pruned to black-box attacks, characterizing a new layer of safety.

In a world that is still hesitant to adopt Generative AI derived from LFMs, in 2025 the market place is awakening to a competitive landscape where leading companies such as IBM with their Granite 3.0 model - commercialising their SFMs to worldwide clients, is also competing against emerging unicorns in the SFMs market from the United States (Liquid AI, Stability AI), Canada (Cohere), France (Hugging Face, Mistral), Israel (AI21Labs), and Singapore (CarbonCopies AI). The most radical entry into this landscape has recently been DeepSeek, an open-source AI company from China, smashing the development paradigms of OpenAI, Anthropic, and NVIDIA, who have built their market dominance on computational power, model parameters, and singular model inference system.

It is very possible that these market events will alleviate the pressure over society and business to adopt LFMs as the only avenue to achieve AI progress. It may well be that some developments of LFMs are discarded or decelerated as governments and industries shift the attention to what these SFMs can offer with a wider scope for AI Safety, and with additional benefits - improved sustainability and energy efficiency as well as financial viability for industry adoption, that seem to be better aligned with government objectives.

2. International proactive steps addressing the fundamental threats to society derived from the misuse of Generative AI

2.1 Misinformation and deepfakes pivoting towards widespread malicious use

AI-produced misinformation and deepfakes have been identified as technological tools used by bad actors and political adversaries in the meddling of democratic processes, spreading propaganda, adversarial attacks and false information. As such, some governments worldwide took significant steps in 2024 to address their growing threat, particularly in elections and public safety.

There is also further pressure on the upkeep of good relations among partner nations and members of international organisations with whom diplomatic channels must be kept open, when the very nature of sophisticated disinformation campaigns often originates from foreign actors.

United States

- *Deepfake Legislation:* The Federal Trade Commission (FTC) drafted rules against the creation and distribution of malicious deepfakes, particularly those aimed at defrauding through the use of impersonation of individuals. Federal Communications Commission (FCC) ruled that calls made with AI-generated voices are "artificial" under the Telephone Consumer Protection Act (TCPA). Several states also introduced legislation to criminalise deepfakes used for political manipulation or harassment. U.S. Congress introduced several bipartisan bills to counter disinformation and misinformation with deepfakes, and enacted TAKE IT DOWN Act in May 2025 which criminalizes the nonconsensual publication of intimate images.
- *Public Awareness Campaigns:* Government agencies launched public awareness campaigns to educate citizens about the dangers of deepfakes and misinformation, encouraging critical media literacy and fact-checking.
- *Election Security Measures:* Increased efforts were made to secure election systems and combat foreign interference, including measures to detect and counter disinformation campaigns.

United Kingdom
The legislation and policy efforts in the U.K. started in 2023 when a landmark bill for online safety was passed at the end of this year including provisions to hold social media platforms accountable for tackling harmful content, including misinformation and deepfakes. The bill required platforms to proactively identify and remove illegal content and to implement measures to reduce the spread of harmful content.

In early 2024, the National Security Bill was passed in an attempt to strengthen the UK's ability to counter foreign interference and disinformation campaigns, including those using deepfakes, providing law enforcement and intelligence agencies with new tools to investigate and disrupt such activities. In 2025 there is a proposal for a specific *Deepfake Legislation* to criminalize the creation and distribution of malicious deepfakes, particularly those aimed at causing harm or spreading disinformation.

Government Initiatives towards the fight against counter-disinformation have also seen the creation of a specific unit to actively monitor and respond to disinformation campaigns, including those using deepfakes. This counter-disinformation unit works with social media platforms to identify and remove false content and to raise public awareness about the threat of misinformation. The UK government has invested in research and development to improve deepfake detection technologies and develop tools to counter the spread of misinformation as well as to raise public awareness to educate citizens about the dangers of deepfakes and misinformation, encouraging critical media literacy and fact-checking. A strong advocate of collaborative approaches to solving problems, the UK government has also been working closely with tech companies to develop industry standards and best practices for tackling misinformation and deepfakes as well as is actively involved in international efforts to address the global threat of misinformation and deepfakes, collaborating with allies and partners to share information and coordinate responses.

As freedom of speech represents a stronghold and a pillar of society in democratic nations, the road to assure online safety while preserving the right to express personal opinions with respect to others who may hold opposing ones represents an enormous challenge in a nation of approximately 56.2 million social media users (Source: DataPortal/We Are Media January 2024). The rapid advancement of deepfake technology requires ongoing adaptation and innovation in detection and countermeasures, forcing the British government to use not legislation, but something new in the hands of legislators: better technologies to fight crime and abuse online.

European Union (EU)

The EU strengthened its *Code of Practice on Disinformation*, requiring online platforms to take greater responsibility for tackling misinformation and deepfakes. The Digital Services Act (DSA) came into force, imposing stricter rules on online platforms to address illegal content, including deepfakes and misinformation. This further continued the objectives sought with the AI Act, which entered into force across all 27 EU member states in August 2024. GPAI had particular input into the AI Act's provisions on deepfakes, in Article 50.2 (on transparency). This article requires that providers of Generative AI systems 'shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated'.

Where it comes to the EU it is important to note that, while the EU AI Act provides an overarching framework, some countries are also implementing their own national measures. In addition to its Network Enforcement Act (NetzDG) of 2023, which included stricter regulations for social media platforms to address illegal content, including hate speech and disinformation, in July 2024, the German Bundesrat (Federal Council) published a draft law specifically targeting deepfakes and proposing to criminalize the creation and distribution of deepfakes that violate personal rights, with penalties of up to two years in prison or a fine; Spain, which in 2023 introduced a draft bill specifically targeting AI-generated images and voices, including deepfakes, and Italy, which is introducing in 2024 on artificial intelligence that includes provisions related to deepfakes and misinformation addressing transparency, accountability, and protection of fundamental rights, are still leveraging from existing legislations around defamation, right to privacy, consumer protection codes, as well as data and intellectual property laws.

Other EU countries are moving more cautiously towards misinformation and deep fakes legislations, opting to deploy softer approaches such as citizens awareness campaigns or addressing deepfakes abuse via existing laws against defamation, fraud, and privacy violations (The Netherlands).

Brazil

In recent years, Brazil has emerged as a compelling case study in the weaponization of AI-driven disinformation. The country's intense political polarization and high social media usage have created fertile ground for the malicious deployment of synthetic audiovisual content. Deepfakes are increasingly used to spread politically motivated disinformation, perpetrate financial scams, damage personal reputations, and erode public trust in institutions. This rapid spread of manipulated content poses a serious challenge, undermining the integrity of democratic institutions and electoral processes.

Brazilian authorities have responded with decisive judicial and legislative measures. For instance, the Superior Electoral Court (TSE) issued Resolution 23.732/2024, banning the use of deepfakes in political campaigns and mandating that all AI-generated content be clearly labelled. Violators now face penalties ranging from fines to candidate disqualification. In parallel, the Supreme Federal Court (STF) has enforced strict compliance measures for digital platforms such as Meta and Google. These measures require the rapid removal of disinformation, hate speech, and deepfakes. Notably, in 2024, the STF temporarily suspended Platform X (formerly Twitter) for non-compliance, lifting the ban only after fines were imposed. Justice Alexandre de Moraes, a central figure in Brazil's electoral governance, has underscored the existential risks of unregulated AI and advocated for expanded judicial oversight to prevent what he describes as "human collective disempowerment." However, some critics argue that these judicial measures risk infringing on free speech and may constitute overreach.

Legislative efforts further complement these judicial actions. In December 2024, the Senate approved Bill No. 2,338/2023—a landmark bill establishing risk-based regulations for AI systems. The bill categorizes AI tools by threat level and mandates transparency, impact assessments, and safeguards against discrimination, while also granting individuals the right to contest AI-driven decisions. Additional proposals, such as the Fake News Bill (PL 2630/2020), seek to tighten rules for political advertising and hold digital platforms accountable for amplifying disinformation.

Beyond legal frameworks, Brazil has launched complementary initiatives to combat AI misuse. The government established a Digital Disinformation Combat Center to monitor synthetic media and coordinate rapid responses during elections. Updates to the National AI Strategy prioritize ethical development, worker protections, and public-sector innovation. This is exemplified by projects like Justice 4.0, which uses AI to optimize court workflows. International partnerships—such as those with Stanford University's Deliberative Democracy Lab—are also being forged to integrate global insights into national policies.

In sum, the malicious use of misinformation and deepfakes in Brazil has become a catalyst for a broader, nationwide effort to regulate artificial intelligence. The convergence of regulatory actions by the government, judicial interventions, and vigilant media and civil society initiatives reflects a determined response to the multifaceted challenges posed by synthetic media. As deepfakes continue to evolve and threaten to blur the lines between reality and fabrication, Brazil's experience serves as a vital case study in balancing the imperatives of innovation and democratic accountability.

China

China implemented stricter regulations on deepfakes and other synthetic media, requiring clear labelling and consent for their creation and use.

India

India introduced guidelines for social media platforms to address misinformation and deepfakes, including mechanisms for users to report and flag false content.

Singapore
Singapore passed a law criminalizing the malicious spread of falsehoods, including deepfakes, with penalties including fines and imprisonment.

2.2 Challenges to be expected in 2025 around misinformation and deepfakes

In 2024 malicious actors increased their use of disinformation and deepfakes in cybercrime, impersonating individuals' voices, video images and messages on social applications used by millions of people daily. While social media users have become less gullible in believing extreme *fakery*, misrepresenting individuals by creating fake personas is increasingly growing as a form of cybercrime to break into people's bank accounts by impersonating either the account holder or the financial institution itself. This shifts the use of *deep fakes* towards an even more dangerous misuse of AI, one that directly affects citizens and vulnerable individuals' lives, as opposed to being solely associated with the widespread of misinformation.

Additionally, social media companies controlling the major platforms such as X and Meta (WhatsApp, Instagram, Facebook) have ceased their detection of deep fakes and misinformation via human fact checking teams. Furthermore, they ceased to implement ethical practices, opting to allow the network of users to self-police in the name of a freedom of speech that very often derives into online mobs that single out individuals for the purpose of verbal abuse, which in some instances has translated to the real world by publishing personal data, such as workplaces and private residences, in order to harangue abuse in real life. Social media platforms, by disregarding to acknowledge how easily and dangerously unrestrained mob mentality can emerge within social networks, allow and condone the inertia of social collectives to silence dissenting voices and to operate within echo chambers that are arbitrary and unpredictable. This attitude and operational mode collides with mainstream practices of crowd monitoring in social gatherings, sporting events, and demonstrations, where security forces are entrusted with tasks such as preventing social unrest and dispersing mobs because individuals alone cannot defend themselves from social disturbance and aggressions. The digital social space should unquestionably be regulated in order to ensure the safety of places where human interactions occur.

Governments must recognize the need to create regulations that demand platform accountability or force their closures in countries where citizen abuse and misinformation infringements occur.

So far, law enforcement agencies have carried out crackdowns in relation to crimes, but in democratic countries, they have traditionally avoided government intervention on internet platforms when it comes to matters that do not necessarily constitute a crime, due to concerns about freedom of expression. Excessive intervention could amount to censorship. Ideally, social media should be a space for free and active discussions, where expressions and ideas are reconsidered and shaped. Moreover, for Intenet Service Providers (ISPs), it would be an undue burden to constantly monitor discussions taking place on their platforms and take actions such as deleting content. For this reason, their liability has been limited. However, considering the growing prevalence of illegal and harmful content on platforms in recent years, it has become necessary to take appropriate measures while ensuring a careful balance between regulation and the protection of freedom of expression.

Ideally, a healthy system would involve the existence of numerous social media platforms, each with its own ISP clearly defining unique rules in its terms of service. It would be desirable for these ISPs to delete statements or suspend accounts of users who violate these rules. Those who dislike a particular platform's rules could simply move elsewhere, and victims who feel uncomfortable due to criticism or other reasons would have the freedom to leave that platform and join another. However, at present, the number of platforms is not that large. As a result, given the limited number of platforms, there is a rationale for the government to step in and regulate the landscape. Instead of directly intervening in expressions, the government should require ISPs to clarify their rules regarding content removal and account suspension.

While many governments have emphasized the importance of public awareness and education to empower citizens to identify and resist misinformation, it is important to also teach users of social media to distinguish the difference between freedom of speech and social abuse, learning to use them with respect to others, or individually suffer the consequences and the weight of the law. If this is enforced, many social abuses may decrease in the long term.

2.3 Algorithmic bias in AI-based systems: How AI agents emerge as a controversial approach to inference automation

Algorithmic bias in AI-based systems poses significant challenges, particularly when it comes to the transparency of advanced AI-based models. As AI technology progresses, a crucial distinction emerges between general AI-based systems and autonomous rational agents. While many AI-based systems make inferences and decisions based on programmed algorithms, only autonomous rational agents are capable of truly independent, autonomic action. This distinction becomes critical when examining the explainability of AI decision-making processes. The increasing complexity of these systems, especially in the case of autonomous agents, makes it progressively more difficult to identify and elucidate the reasoning behind their inferences and decisions. This opacity raises important questions about accountability, fairness, and the potential for unintended consequences in AI-driven outcomes At Davos 2025, both Nobel Prize winner Dr. Demis Hassabis and ACM A.M. Turing Award Winner Dr. Yoshua Bengio expressed their concerns about the use of *agentic* AI in developing advanced AI systems.

Dr. Bengio highlights the inherent unpredictability of highly complex AI-based systems and argues that as AI technology becomes more sophisticated and autonomous, it becomes harder to anticipate and control its behaviour, potentially leading to unintended consequences. He stresses the need for responsible AI development, with a focus on safety and alignment with human values and advocates for increased research into AI safety and ethics, as well as international collaboration to ensure the safe and beneficial development of AI. Dr. Bengio has been vocal about the need for AI governance and regulation to mitigate potential risks and believes that proactive measures are necessary to prevent harm and ensure that AI benefits humanity. Dr. Bengio calls for increased public awareness, significant investment in AI safety research, and the creation of democratic, decentralized coalitions for AI development to ensure that the technology serves the common good rather than narrow interest.

Dr. Demis Hassabis, CEO of Google DeepMind, stresses the need for cautionary approaches to *agentic* AI within the ambivalent context in which current developments present both the immense potential of AI while emphasizing the need to develop AI with safety and ethics as top priorities that are aligned with human values and goals. Hassabis believes that AI should be a tool for good and that its development should be guided by human needs and aspirations. Hassabis has expressed openness to AI regulation and the need for international cooperation to ensure the responsible development and deployment of AI.

What makes Bengio's and Hassabis' concerns a real issue in 2025 is that the paradigm upon which artificial general intelligence has been developed replicates human intelligence, which is biological and therefore seeks to exist, making agentic AI carry a DNA of survival at all costs, and therefore a concern that, if alerted of the danger of being interrupted, it may find ways to obstruct such action.

2.4 Understanding agents technology and the risks that they represent

The revamp of autonomous agents technology in the context of Generative AI (GAI) represents a significant paradigm shift in AI research and applications. These advanced systems, often called *agentic AI*, have evolved from simple rule-based chatbots to complex, adaptive entities capable of independent decision-making and task execution. Large language models (LLMs) and other advanced AI-based technologies allow autonomous agents to anticipate needs, develop strategies, maintain context, and orchestrate multiple tools and resources.

Autonomous agents are AI-based systems designed to operate independently, executing tasks through sensing, processing, and adaptive decision-making. These agents employ frameworks like the Belief-Desire-Intention (BDI) model, which formalizes goal-directed behaviour using logical reasoning about beliefs, desires, and intentions. They integrate machine learning and natural language processing (NLP) to analyse environments, predict outcomes, and dynamically adjust strategies. Applications span automated customer service, supply chain optimization, and real-time data analysis, with capabilities like tool integration (APIs, databases) and multimodal perception (text, audio) enabling complex problem-solving. Prof. Wooldridge's foundational work[2] highlights their capacity for teamwork and communication in multi-agent systems, while Prof. Sierra's research emphasizes their role in computational trust and artificial social ecosystems[3].

LLMs extend autonomous agents by enhancing natural language kind-of understanding and generative reasoning. LLMs enable agents to parse unstructured data, simulate *human-like* interactions, and refine plans through fine-tuning domain-specific datasets. For instance, architectures like NatBDI combine BDI reasoning with NLP-driven belief updates and plan libraries, allowing agents to process observations in natural language and execute context-aware actions.

[2] Wooldridge M, Jennings NR. Intelligent agents: theory and practice. The Knowledge Engineering Review. 1995;10(2):115-152. doi:10.1017/S0269888900008122

[3] Montes, N., Osman,N., Sierra, C. & Slavkovik,M. Value Engineering for Autonomous Agents. CoRR abs/2302.08759 (2023)

Architectures that use LLMs specifically as inference engines for agent reasoning, such as ReAct[4] proposed by Yao, Zhao et al., have recently emerged. This integration supports scalable social simulations and adaptive decision-making in dynamic environments, though challenges persist in aligning LLM outputs with ethical frameworks. Prof. Dignum's principles of responsible AI underscore the need for transparency in such hybrid systems, ensuring agents balance autonomy with accountability. Together, these advancements position LLM-augmented agents as transformative tools for modelling complex sociotechnical systems.

A critical concern is the possibility of agents making *harmful* inferences or decisions based on incomplete or misinterpreted information. For example, an autonomous agent might recommend *inappropriate* treatments in healthcare applications due to misunderstanding context or failing to consider all relevant factors. Additionally, using LLM-augmented agents in sensitive domains like national security or financial markets could lead to unintended consequences if their decision-making processes lack transparency or human oversight. To mitigate these risks, it is crucial to implement robust ethical guidelines, regular audits, and human-in-the-loop systems to ensure responsible development and deployment of autonomous agent technologies.

The development and deployment of autonomous agents, particularly those augmented by LLMs, raise significant concerns regarding alignment with EU values of ethics, legal compliance, security, and privacy (ELSEC[5]). As Taddeo *et al*[6]. argue these AI-based systems must be designed with "ethics-by-design" principles to uphold fundamental rights and societal values. A critical aspect of this alignment is the explainability of agent decision-making processes. Arrieta et al. (2020) emphasize in Information Fusion that explainable AI (XAI) is essential for building trust, facilitating audits, and enabling meaningful human oversight. In the context of autonomous agents, this translates to the need for transparent reasoning mechanisms that can elucidate how beliefs, desires, and intentions lead to specific actions. Doshi-Velez and Kim[7] that such explainability should be evaluated in the context of the specific task and end-user, suggesting a nuanced approach to agent transparency. Furthermore, as Dignum *et al*[8]*.,* responsible AI development requires technical solutions and consideration of societal impact and ethical frameworks. This multifaceted approach to explainability and ethical alignment is crucial for ensuring that autonomous agents, especially those leveraging powerful LLMs, can be safely and responsibly integrated into critical EU sectors while maintaining public trust and adhering to ELSEC values.

2.5 Geopolitics influencing AI development and its economic and social impacts

Geopolitics will play a significant role in shaping the development of Advanced AI in 2025 and beyond. In the United States, the new administration has immediately withdrawn participation in and support

[4] ReAct: Synergizing Reasoning and Acting in Language Models. https://react-lm.github.io

[5] ELSEC stans for Ethics, legal, Socioeconomic and Cultural uses of AI-Based applications.

[6] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[7] Doshi-Velez, F., & Kim, B. (2017). Towards A Rigorous Science of Interpretable Machine Learning. arXiv preprint arXiv:1702.08608.

[8] Dignum, V., Baldoni, M., Baroglio, C. et al. (2018). Ethics by Design: necessity or curse?. AI and Society, 33, 401–419.

for collaborative efforts at major international alliances around climate change and world health by stepping out of the Paris Accord and the World Health Organisation. With the U.S. AISI hosted within the U.S. Department of Commerce, and the emphasis for independence and extreme competitiveness that this new government seeks, it is foreseeable that unethical and unsafe approaches to advanced AI-based systems may be condoned in the name of unbound capitalism without regulatory restraint. With the United States and China locked in a race for AI dominance, both viewing Advanced AI as a strategic asset with implications for economic growth, military power, and national security, AI-based technology development may be even further accelerated at the expense of cutting corners on safety and ethics.

The potential for Advanced AI to revolutionize warfare is driving investment in AI for military applications, raising concerns about autonomous weapons systems and the potential for AI-driven escalation that will by-pass previous international agreements about ethics in warfare arms development. The current war in Ukraine, the conflict in Yemen, the Ethiopian Civil War, and the Syrian Civil War together with smaller-scale conflicts occurring in countries like Afghanistan, Myanmar, Nigeria, Somalia, and Sudan, often involving ethnic or religious tensions, political instability, or competition for resources, may be difficult to be de-escalated in 2025 if other tensions related to destabilization of the Middle East, maritime security for shipping cargoes, especially those transporting crude oil barrels and energy supplies, involve AI developing nations who may increase espionage and attempts to sabotage rival nations' progress, potentially hindering overall development and creating instability by deploying rogue AIs. At the EU level some think-thanks start to discuss issues like how “The race to control disruptive technologies will intensify”[9].

The International Network of AI Safety Institutes Mission Statement specifically underwrites the need for global inclusion of efforts towards AI safety, planning to “actively engage countries, partners, and stakeholders in all regions of the world and at all levels of development by sharing information and technical tools in an accessible and collaborative manner, where appropriate”, promoting not only AI Safety but also security, inclusivity, and trust as vital components in addressing risks, driving responsible innovation, and expanding access to the benefits of AI worldwide.

[9] https://commission.europa.eu/document/download/5bb2881f-9e29-42f2-8b77-8739b19d047c_en?filename=2024_Niinisto-report_Book_VF.pdf

3. Potential avenues for the improvement in awareness of advanced AI by business and society

Business and society are not yet fully acquainted with what artificial intelligence is, and there are many degrees of acceptance of its benefits depending on which nations and which active development of automatization systems, data protection laws, and general pessimism regarding unemployment and the displacement of human jobs because of artificial intelligence.

3.1 Societal adoption of artificial intelligence

The general public, thanks to media coverage and popular applications like ChatGPT, seems to grow awareness in regard to its usefulness, yet remains unaware of the nuances of advanced AI, its capabilities, and limitations. In spite of data privacy regulation, digital users still over-share personal data with technology companies, unaware that they have the right to decline certain functionalities. In this context, it is even more important to create a clear understanding of the right to put barriers to AI systems that gather information and proprietary content from digital users without disclosing this fact or what they intend to do with it. An example of this is Instagram - a social media property of AI developer Meta, and how creative content posted by users is indirectly owned by Meta as their new terms and conditions state, which also do not disclose the use that they make of it. The suspicion is that such images are used to train Meta's AI-based systems, something that other companies such as Google have also done in the past, for example by acquiring human verification company Captcha - originally developed by Carnegie Mellon imagen recognition researchers, Google has been able to use human image identification to train the computer vision systems of another of its acquired companies, WAYMO, who develops autonomous driving technology. The first version of ChatGPT was released to the public still in training with the intent of using early adopters as testers and trainers of its AI-based system, an unprecedented event that it is not allowed to other technology sectors who are required to release products and services to the wide public fully tested and adherent to safety standards. It is important to note that, while it is standard practice in software to release products under a user-testing mode denominated a "minimum-viable-product" or MVP, this is released to a controlled group of individuals, never on a global scale of millions as it was the case with OpenAI.

If AI continues to be released to people unbound, unrestricted, unproven and unapologetically malfunctioning, with obscure terms and conditions that users do not understand, and preying on their data privacy, governments will face societal backlash if any issue concerning AI is proven to originate from this neglect in creating guardrails for when digital users are indirectly turned into easy prey for companies.

3.2 Adoption of AI-based applications in businesses

Many businesses are exploring AI-based applications, but the level of understanding varies greatly as the challenges around lack of skilled talent, data limitations and integration complexities hinder faster adoption. Since the majority of businesses are represented by small to medium enterprises in most developed countries, it is fair to say that adoption of artificial intelligence in the business world remains in the early stages. Some industries are actively investing in AI research and development, for example financial services, logistics, transportation, automotive and manufacturing, while other sectors are still in the early stages of exploring potential use cases. This is partly due to the slow adoption of data governance practices required for basic data analytics operations, the cost of upgrading infrastructures and legacy systems, and the unclear return on investment for operational efficiency via artificial intelligence, hence full-scale integration is relatively rare. Businesses only started to transform their assets into digital attributes in the 2010s. Estimates suggest that even in advanced economies, at least 20-30% of operations could be considered primarily analog. Against this backdrop, the full adoption of advanced AI systems remains a complex process because of several dependencies that range from availability of data, computing power and affordable and trustworthy AI-based tools and platforms.

3.2.1 Data availability

AI models require large, high-quality datasets for training and development. Robust data that guarantees provenance and integrity is likely to also be data that can be trusted for its truthfulness. This can only be achieved if businesses gather, process and store data via infrastructure that abides by standards and legitimacy. Many of the data infrastructures in business present combinations of data architectures and storage facilities and paradigms that can facilitate or make it more cumbersome for businesses to fully deploy AI technologies. Data curation, a critical aspect of AI-based systems development, comes with significant costs but is essential for ensuring the quality and reliability of AI models. However, investing in proper data curation can lead to substantial benefits, including improved model performance, reduced risk of bias and errors, and enhanced overall AI workflow efficiency.

3.2.2 Computing power

Both training and inference for advanced AI-based models, especially deep learning models, demand substantial computing resources. Consequently, businesses require access to high-performance hardware, which can be obtained either on-premises or via cloud-based services. This explains the direct effect of Cloud computing on adoption of advanced AI technologies. Cloud providers like AWS, Google Cloud, and Microsoft Azure offer affordable access to powerful computing resources which facilitates that small and medium enterprises leverage from these services to train and deploy AI models without significant upfront investments in hardware. Still, there is an additional element of interdependence where it comes to Cloud computing services and AI adoption: cybersecurity.

3.2.3 AI-based tools and platforms

Building and maintaining robust cybersecurity infrastructure in-house is expensive. Cloud providers offer this as part of their services, saving SMEs significant upfront and ongoing costs. Cloud providers have dedicated teams of security experts who manage and update their security infrastructure, ensuring a higher level of protection than most SMEs could achieve on their own, especially when trying to hire this human capital directly is complicated. Cybersecurity skills are highly in demand and their salaries are unaffordable for many businesses. Building highly efficient cybersecurity teams takes years and the security requirements increase in sophistication as cyber crime continues to leverage increasingly from new technologies. It is a requirement that has exponentially become unattainable for SMEs. In addition to expert human capital, cloud-based security must scale in synch with the SME's needs, adapting to changes in data volume and usage patterns without requiring manual adjustments. This explains why major cybersecurity companies have partnered with cloud computing infrastructure platforms and developed AI-based solutions such as Zero Trust platforms. These security platforms are based on the paradigm that nothing should be trusted, and everything affecting data should be checked.

Additionally, cloud providers often help SMEs meet regulatory compliance requirements for data security, such as GDPR or Health Insurance Portability and Accountability Act (HIPAA), by providing built-in security features and certifications. This is an important issue that governments must not overlook and leave to the marketplace to deal with via its own resources. If companies fully outsource all aspects of data to third parties, the convenience will debilitate the need for internal resources that will guard against breaches in data governance. Balancing this must be part of the policy discussions about direct and indirect responsibility over data governance related to assuring artificial intelligence.

Governments must take an active role in driving the awareness of advanced artificial intelligence amongst both business and society, fundamentally for safety reasons - educating the population into the safe use of AI and the acquisition of skills that will prevent them to fall for cybercrime or digital abuse of their data or mental vulnerabilities, and to encourage the optimization of businesses, the reduction of costly errors in healthcare, and the need to strengthen the innovation and competitiveness of national businesses.

The AI-Factories initiative aligns well with the concept of cost-effective cybersecurity infrastructure for SMEs. The European Commission's €1.5 billion investment in AI Factories[10] aims to provide centralized access points for AI development and validation, including infrastructure and services.

Adoption of artificial intelligence in its most advanced stages of development requires adequate infrastructure and incentives towards digital transformation and trustworthy data governance that allows businesses to invest in this type of innovation and direct government support of this will contribute to guarantee safety and trustworthiness.

[10] https://digital-strategy.ec.europa.eu/en/policies/ai-factories

4. Advancing AI safety: International efforts to assure the trustworthiness of AI ensuring both Innovation and societal welfare.

4.1 Targeted aims: International cooperation via national approaches

Both the Hiroshima Process and the Bletchley Declaration aimed to foster and sustain international cooperation and the creation of the International Network of AI Safety Institutes (AISI) seeks to align on priority work areas that can advance the global collaboration on AI safety, still allowing local national approaches. In practice, AISIs are converging into adopting common practices wherever possible, however, certain type of model validations will depend on a variety of conditionings such as access to models, financing, development of AI awareness in government in order to prioritise support for validation activities, and, inevitably, the flexibilities or impossibilities of certain market conditions, policies, regulations and industry acceptance of AI.

The first steps towards this goal began In May of 2024, when the United Kingdom and the United States AISIs set up similar sounding missions and goals for the year, yet deploying efforts quite differently. The UK AISI, part of the UK government, set out to assemble the right technical expertise in-house, hiring AI scientists with skills and experience with advanced AI models. This headcount of highly specialized human capital has run tests on Frontier AI systems aiming to understand the risks that those systems might pose from a technical perspective, not a regulatory one, developing tests across a wide range of domains in order to audit the potential misuse of AI systems, specifically focusing on attacks, whether chemical, biological, or cybernetic, targeting national infrastructures and the population. In addition, analysing the societal impact when abuses affect users, or the potential risks arising from such systems' autonomies that arise from the autonomies of such systems when they set and execute goals without human control or final authorisation, have also been added to their testing objectives. After the Bletchley Park Summit, UK AISI began recruiting and growing its headcount very publicly, posting job opportunities on their website and appearing in as many public events as possible.

As highlighted at the first meeting of the International Network of AISI[11] in San Francisco in November 2024, the UK government hosted a conference on how AI developers can put into practice the commitments made at the AI Seoul Summit. The event was co-organized with the Centre for the Governance of AI and led by the UK's AISI and aimed to give AI companies and researchers a clear focus on how they can bolster their AI safety plans to accelerate the design and implementation of frontier AI safety frameworks. Although the Frontier Models development is forcing AISIs to amplify their efforts towards these AI systems, which present grave challenges for humanity in the next five to ten years, Generative AI must be developed and commercialized within controlled frameworks because it is now a commercially available technology rather than a research endeavour, within the confines of a laboratory. In February 2025, the institute has been renamed as “AI Security Insitute.”

[11] The International Network of AI Safety Institutes has a diverse group of initial members from around the globe. The countries and entities that were part of the network as of November 2024 include: Australia, Canada, European Union, France, Japan, Kenya, Republic of Korea, Singapore, United Kingdom, and the United States.

The International Network of AISI have committed an USD 11 million to constantly measure its impact on society and business, which is both beneficial and negative, and specifically outline defined measures to fight the synthetic content risks, which increasingly challenge the trustworthiness of models, and the prioritization of creating industry standards when Generative AI sold at scale and on a global basis by technology vendors.

4.2 Technical validation: A best-efforts approach with limitations

In January 2024, the UK AISI was provided with access to one of the big Frontier Models before it was released so that they could run their tests and present feedback to the developer. The strategic value of this approach is that it allows the technical testers to become familiar with the model's inner workings of the model, expanding the scope of the additional tests that should be run for specific vulnerabilities of the system, which could derive further specific risks. In return, the UK AISI was able to reveal to the AI developers what specific issues make the UK government uncomfortable when it comes to National Security, becoming not only a qualified mediator between both parties, but also one that performs empirical approaches rather than legal approaches to validation. The long-term objective is to improve the discussions around aiming for standards and for the internationalization of technical testing efforts beyond the United Kingdom.

The UK AISI and its US counterpart developed methods to assess and mitigate risks of advanced AI systems, created benchmarks and evaluation tools for AI models, established safety guidelines for AI applications and conducted joint tests together.

Together, the United Kingdom and the United States, have been able to successfully demonstrate in 2024 that this approach is the most appropriate and rewarding in addressing the most pressing challenges of Generative AI models, which are primordially technical, but limited to the AI developers' willingness to continue to share models in training. It is thus not guaranteed that they continue to do so if the marketplace becomes increasingly competitive and government administrations allow for AI developers to steer away from the good will frameworks of international cooperation.

Around these joint efforts, there are other national approaches worth mentioning.

The Center for AI Standards and Innovation (CAISI)

The U.S. AI Safety Institute (AISI) was established in 2023 with the Executive Order 14110 on the Safe, Secure, and Trustworthy Development and Use of AI. In June 2025, its has been renamed as The Center for AI Standards and Innovation (CAISI). The CAISI operates under the Department of Commerce and is nested within the National Institute of Standards and Technology (NIST), which is responsible to develop and monitor science-based technological standards. CAISI's objectives are to 1) Work with NIST organizations to develop guidelines and best practices to measure and improve the security of AI systems, and work with NIST staff to assist industry to develop voluntary standards, 2) Establish voluntary agreements with private sector AI developers and evaluators, and lead unclassified evaluations of AI capabilities that may pose risks to national security, 3) Lead evaluations and assessments of capabilities of U.S. and adversary AI systems, 4) Lead evaluations and assessments of potential security vulnerabilities and malign foreign influence arising from use of adversaries' AI systems, 5) Coordinate with other federal agencies and entities to develop evaluation methods, as well as conduct evaluations and assessments, and 6) Represent U.S. interests internationally CAISI's redefined mission prioritizes standards development, evaluation, and voluntary guidance over regulatory mandates.

CAISI plans to meet its objectives by performing and coordinating technical research to improve or create guidelines and evaluations.

Although the CAISI has an expansive remit, it also faces certain challenges which may impact its ability to meet the expectations.

- **Authority:** while the CAISI conducts advanced research, this work is dependent on voluntary cooperation of the private sector and access to proprietary advanced models. The outcomes of the CAISI are voluntary guidance, rather than regulations or binding guidelines.
- **Commercial interests:** the need for close ongoing cooperation with the private sector may result in prioritization of commercial concerns or interests over societal ones. Additionally, a strictly technical assessment may give false assurance about systems which are inherently deployed in social contexts.
- **Resources:** funding and personnel of the CAISI is dependent on annual budget allocation, making it harder to plan long-term projects.

To foster domestic collaboration between different actors, and also develop more AI safety resources, former U.S. AISI had launched a Consortium (AISIC) which brings together 280 organizations. The initial focus is to establish standardized evaluation protocols and benchmarks to assess risks across different domains. These protocols are then to be used in both pre and post testing of advanced AI models by CAISI.

Given the rapid evolution of advanced AI systems, the need for global governance harmonization, and the resource constraints on AISIs at large, global cooperation is necessary. So far, the U.S established bilateral partnerships with the United Kingdom, Canada, EU and Japan to deepen collaboration. U.S also led the effort to establish a global network of AISIs to minimize duplication of resources and also bring together global technical expertise. The global network held the inaugural meeting in San Francisco in November 2024. The global network is envisioned to advance both the science of AI safety and enable cooperation on research, best practices, and evaluation. Initial members of this network are Australia, Canada, EU, France, Japan, Kenya, the Republic of Korea, Singapore, the United Kingdom, and the United States. While the U.S. led the coordination of international network in its first year, the role can be rotated between other member countries.

Former U.S. AISI had also established a TRAINS Taskforce to specifically focus on AI risks related to radiological, nuclear, chemical and biological security, cybersecurity, critical infrastructure and conventional military capabilities. The taskforce brings together experts from Commerce, Defence, Energy (with its National Laboratories), Homeland Security, National Security Agency, and National Institutes of Health.

By January 2025, former U.S. AISI has released 2 pre-deployment reports for advanced models. Joint pre-deployment evaluation of Anthropic's Claude 3.5 Sonnet by U.S. AISI and UK AISI. Both institutes ran separate but complementary tests across four domains: (1) biological capabilities, (2) cyber capabilities, (3) software and AI development, and (4) safeguard efficacy. The report notes that "tests were conducted in a limited time period with finite resources" and that "partial assessment of model capabilities at a particular point in time and relies on evolving evaluation method."

The US and UK AISIs also conducted joint pre-deployment evaluation of OpenAI's o1 across three domains: (1) cyber capabilities, (2) biological capabilities, (3) and software and AI development. The report does not explain why OpenAI's testing did not include "safeguard efficacy" as a domain.

Japan AI Safety Institute (J-AISI)[12]

Japan AI Safety Institute (J-AISI) was officially launched on February 14, 2024, following the United Kingdom and the United States. Based on the "Integrated Innovation Strategy 2024" (approved by the Cabinet on June 4, 2024), J-AISI has been positioned within the Information-technology Promotion Agency (IPA) as a central institution for AI safety. It is an organization formed with the cooperation of 10 relevant ministries and 5 related organizations. Its roles are to:

1) support the government by conducting surveys on AI safety, examining evaluation methods, and creating standards,
2) work as a hub for AI safety in Japan by consolidating the latest information in industry & academia, and promote collaboration among related companies and organizations,
3) collaborate with AI safety-related organizations.

It adopts a flexible approach to addressing AI-related issues, including social impact, AI systems, data, governance, and content, while aligning with global trends. To date, it has undertaken initiatives in the following areas:

1) Released the Japan-U.S. Crosswalk in collaboration with the U.S. NIST, which highlights the interrelationship between the U.S. NIST AI Risk Management Framework (RMF) and Japan's AI Guidelines for Business (GfB).
2) Released the "Guide to Evaluation Perspectives on AI Safety," which outlines a foundational approach for assessing the safety of AI systems. This guide serves as a reference for operators developing and providing AI and includes:
   Risks and evaluation items considered in safety assessments.
   Conductors and timing of evaluations.
   An overview of evaluation methodologies.
3) Released the "Guide to Red Teaming Methodology on AI Safety," which offers essential considerations for applying the red teaming method—an approach used to evaluate AI system safety. This guide serves as a reference for operators and includes recommendations on conducting structures, timing, planning, methods, and improvement plans for safety assessments.

J-AISI has also participated in international collaborations, including contributing to discussions at AI-related summits and engaging in dialogues with global business operators and organizations.

[12] https://aisi.go.jp/assets/pdf/20250106_AISI_en.pdf

Agencia Española de Supervisión de la Inteligencia Artificial

The Agencia Española de Supervisión de la Inteligencia Artificial (Spanish Agency for the Supervision of Artificial Intelligence) (AESIA[13]) is an autonomous agency established on September 2, 2023, under the Spanish Department of Digital Transformation. Its primary role is to oversee, counsel, and promote the responsible use and development of artificial intelligence (AI) systems in Spain. This includes responsibilities for inspection, verification, and enforcement to minimise potential risks associated with AI technologies.

Key Functions of AESIA:

- Oversight and Counselling: AESIA provides guidance on the ethical and safe application of AI-based technologies.
- Awareness and Training: The agency conducts training programs to educate stakeholders on AI development best practices.
- Inspection and Verification: AESIA is tasked with ensuring compliance with AI-based system regulations.
- Sanctioning Authority: The agency can impose penalties for non-compliance with established AI guidelines.

The formation of AESIA is part of a broader initiative by the Spanish government to enhance the governance of AI, which began with the creation of the Secretariat of State for Digitalization and Artificial Intelligence in 2020. The establishment of this agency aligns with Spain's National Artificial Intelligence Strategy, which aims to foster ethical standards in AI-based applications development.

[13] https://mpt.gob.es/politica-territorial/desconcentracion-sector-publico-institucional-estatal/determinacion-sede-AESIA.html

5. Significant differences in how AI safety is approached reflecting varying levels of regulatory frameworks and cultural priorities[14].

5.1 The various regulatory frameworks: Hard law regulation (EU) vs. soft law regulations (U.S., UK, Japan)[15]

The European Union's Approach to AI Safety
The EU AI Office's mandate can be found in the EU AI Act which aims to protect safety and fundamental rights of people. The Office is charged with implementing the AI Act and supporting the respective governance bodies in the Member states. For general-purpose AI systems, the Office enforces the rules and is also charged with establishing code of practice and future evaluations[16]. The AI Act enables the Office to conduct evaluations of general-purpose AI models, request information and measures from model providers, and apply sanctions where necessary. Such a legal foundation gives it a more solid foundation and makes it much harder to roll back, dismantle or even reduce the powers and authorities of the entity. An "AI Safety" unit within the Office is tasked with developing tools, methodologies and benchmarks for evaluating capabilities of general-purpose AI models, and classifying models with systemic risks so that relevant obligations can be applied. To complement the work of the Office, collaboration is established with the Scientific Panel of independent experts, and the Advisory Forum, representing a variety of stakeholders. In addition to the EU AI Act, other product and liability legislation also apply in vertical domains.

An emerging concern is that as safety institutes proliferate, companies may choose only to cooperate and provide access to models where they are legally obliged to do so (i.e. either in their domestic jurisdiction or where there are hard law requirements such as in EU AI Act). Coordination and information-sharing between AISIs will be key to ensuring safety, advancing understanding and science of safety, and collectively building capacity.

The United States' Approach to AI Safety
U.S. AISI was initially established by executive order from the President Biden, and not by congressional action, and Trump Administration renamed the organization as the Center for AI Standards and Innovation (CAISI) in June 2025. Therefore, change in administrations can impact the continuity and mandate of the AISI in the U.S. The entity was focused on "advancing the science of AI safety," and does not have any regulatory powers. However, this does not mean that AI safety at large is not regulated. Existing regulatory agencies can investigate companies and products which fall under their authority (The Federal Trade Commission: consumer safety from AI-based products, U.S. Food and

[14] 'Safety' means different things in different national and historical contexts. Not only there seems to be no consensus upon an agreed definition of 'safety' but the term across languages and cultures encompasses /omits different things.

[15] Although US, UK and Japan are taking soft law approach, that doesn't mean they are loose in regulations. Japan is applying or amending existing laws in relations to AI (i.e. combinations of soft law AI Guidelines for Business + amendments to existing laws). U.S. is too. Also, UK and Japan are taking similar approaches.

[16] Later on, AI Office was tasked with coordinating the formulation of the General Purpose AI Code of Practice with the companies that may later be subjected to the requirements of the AI Act.

Drug Administration: patient safety of AI-driven medical products, National Highway Traffic Safety Administration: automated vehicles, etc).

The United Kingdom's Approach to AI Safety

The United Kingdom is combining a distinct approach to AI safety, aiming for a balance between fostering innovation and mitigating potential risks. Instead of creating a whole new regulatory body or sweeping legislation for AI, the UK is relying on existing regulators to interpret and apply five broad principles to AI within their respective sectors. These principles are:

- Safety, security, and robustness
- Appropriate transparency and explainability
- Fairness
- Accountability and governance
- Contestability and redress

This approach allows for flexibility and avoids one-size-fits-all rules, adapting to the specific context of AI use. Notably, the UK is also prioritizing research into "systemic AI safety," which looks at the broader societal impact of AI, including potential risks like deepfakes, misinformation, and AI system failures in critical sectors. This focus aims to identify and address long-term, widespread risks that could affect society as a whole.

While UK changed the institute's name to AI Security Insitute in Fevruary 2025, the work focus stayed mostly same. UK is also committed to fostering AI innovation, aiming to create a regulatory environment that encourages the development and adoption of AI technologies while ensuring responsible use.

The challenges derived from the principles-based approaches require certain considerations relative to how they are to be enforced. As the principles-based approaches rely on regulators, these may be led to display inconsistencies in the enforcement of principles and show potential gaps in oversight. Furthermore, as AI continues to be developed more rapidly and at global scale, it may be challenging to keep the principles and regulatory framework up-to-date, showing a poor level of adaptability to market shifts. Balancing innovation and the need for standards is a wicked issue that will only derive optimal results if innovation is redefined as a beneficial and powerful way of developing a kind of progress is inclusive of principles from the outset and the concept phase. This is why it is imperative that AI safety is guaranteed by a "safe by design" development mode.

Japan's approach to AI Safety

Japan takes the soft law approach to regulate AI comprehensively. In April 2024, AI Guidelines for Business was adopted by merging the existing three guidelines on AI R&D, Utilization and Governance. In order to move forward with the concrete implementation of principles related to AI, the following matters were pointed out[17]:

- AI use is viewed as a solution to some social challenges, such as decreasing labor caused by a declining birthrate and aging population.
- There is a time lag between formulation and enforcement of laws and the speed and complexity of AI technology development and social implementation.
- Rule-based regulations that stipulate detailed obligations might inhibit innovations.

Thus, it was decided to draw up guidelines on the basis of the goal-based concept that would lead to the achievement of purposes through soft laws without any legally binding force that would encourage

[17] Ministry of Internal Affairs and Communications, Ministry of Economy, Trade and Industry, AI Guidelines for Business Ver1.0

interested parties to make voluntary efforts to reduce societal risks in AI and promote innovations and use of AI.[18]

However, Japan is also amending existing laws on a sector-specific basis, meaning that AI regulations are implemented through a combination of comprehensive soft law guidelines and sector-specific hard laws.

Soft law, which lacks legal binding force, is often misunderstood globally as being unenforceable. However, even without legal binding force, compliance is expected in Japan. The country has a corporate culture in which companies tend to follow directives from the government and regulatory authorities. As a result, Japanese companies have adhered to such guidelines since the previous regulations were introduced.

Notably, Japanese companies are highly sensitive to product safety, as evidenced by their approach to ensuring the safety of their products. Additionally, Japan has a strong culture of social sanctions, making companies highly averse to failures or actions that could damage their reputation.

Despite this, if cases arise—especially involving foreign businesses—where unethical AI development occurs due to non-compliance with guidelines, there is a possibility that such guidelines could be converted into hard law. In May 2025, Japan enacted the Act on the Promotion of Research and Development and the Utilization of Artificial Intelligence-Related Technologies," which promotes adoption of AI by corporations and individuals, while expecting trust-building activities. The Act also legalizes government investigative authority and the obligation for companies to provide information in case of serious incidents.

[18] *Ibid.*

6. Arguments around the balance between AI research funded by governments versus unbound industry-led innovation

Governments play an active role in contributing to the development of economic and social progress via the direct funding of research facilities, the awarding of grants and the creation of regulatory environments that support competitive markets. The context in which Generative AI has been released into the world, industry-led, and by way of imposed disruption, has forced the discussions around how to best support the future development and commercialization of Generative AI towards assessing the best way to combine both government and private funding.

Government funding offers outstanding value in several fronts:

- **Focus on long-term research:** Governments fund research with a longer-term horizon, focusing on fundamental advancements that may not have immediate commercial applications but are crucial for scientific progress and societal benefit because nations seek academic excellence via the contributing value of their research and educational academic centres.
- **Science as a search for truth for the benefit of society:** Government-funded research prioritizes projects that address societal challenges, such as healthcare, environmental sustainability, or accessibility, which may not be as commercially attractive to industry. Science – from Aristotle and Roger Bacon to Albert Einstein and Richard Feynman, must seek the truth, following the inherent human drive to understand the world, always open to questioning our own beliefs and seeking evidence that could challenge any theories. This reinforced the need for rigorous methodology, peer review, and the constant quest for new knowledge that accurately reflects reality, processes and protocols that do not exist as such in commercial innovation, which seeks to solve a problem or fulfil a need gap at all costs, and that in the context of Generative AI has drawn much debate as to what the role of science in the 21st century should be. Some argue that the primary goal of science should be utility, meaning the development of technologies and solutions that benefit society, applying scientific discovery towards endeavours that respond to concrete problems, rather than follow the paths of discovery which sometimes take science into impractical theoreticism. Still, while utility is important, the pursuit of truth provides the foundation for reliable and effective applications of scientific knowledge that should influence how innovation builds reliable products. Government funding must advance transparency, rigour and reproducibility in science.
- **Openness and collaboration:** Government-funded research aims to encourage openness and collaboration, leading to the sharing of knowledge and resources that can benefit the broader AI community.
- **Addressing market failures:** Governments can step in to fund research in areas where the market may fail to invest sufficiently, such as AI safety or the development of public datasets.

Historically, the success of government-led innovation via research facilities has allowed for the development of break-through technologies such as the internet, GPS in cellular telecommunications, and many medical discoveries initially funded by government research. In AI, government funding has played a crucial role in supporting early research and the development of foundational technologies via universities, where breakthroughs in AI development were achieved at U.S. universities (MIT, Berkley, Stanford, Carnegie Mellon), Canada (Toronto), England (Oxford), Switzerland (ETH Zurich), France (Sorbonne, PSL), Japan (Tokyo, Kyoto) and others. Some of these bright AI researchers have ended up founding tech companies (Hassabis co-founding DeepMind; siblings Dario and Daniela Amodei left OpenAI to found Anthropic; Aidan Gomez, Ivan Zhang, and

Nick Frosst, researchers at the University of Toronto and Google Brain, founded Cohere; Clément Delangue, Julien Chaumond, and Thomas Wolf co-founders of Hugging Face; or joining major AI development companies (Yann Le Cunn at Meta; Fei-Fei Lin at Google Cloud AI/ML), successfully putting research towards commercial endeavours.

In parallel to this, and going beyond where government funding does not suffice, industry-led innovation presents very specific dynamics and advantages:

- **Agility and market responsiveness:** Industry-led innovation can be more agile and responsive to market demands, leading to faster development and deployment of AI applications that identify service gaps or real world-needs.
- **Commercialization and scaling:** Industry has the resources and expertise to commercialize and scale AI technologies, bringing them to a wider audience and maximizing their impact, whether they are global leaders with wide geographical and commercial reach or they develop AI under open-source paradigms, which allows the hyper-active global community of AI developers to engage.
- **Competition and efficiency:** Competition within the industry can drive innovation and efficiency, leading to better products and services, especially when companies opt to create competitive attributes based on better products rather than fighting the pricing battles. In addition, companies that create networks and industry clusters with suppliers and partners, are incredibly efficient in taking products to market.
- **Focus on practical applications:** Industry-led innovation tends to focus on developing AI applications with clear practical uses, which can lead to tangible benefits for businesses and consumers.

The movement of prominent AI researchers into industry is a testament to the growing importance and potential of AI. It highlights the increasing connection between research and real-world applications, and underscores the need for continued focus on responsible AI development that benefits society as a whole. The acceleration of AI developments in the last ten years has created a competitive landscape where companies developing commercial innovation seek the brightest talent from academic institutions that work together with other talented individuals with product development skills increasingly focused on human centric values – the DNA of industrial design, and ethical considerations around innovation that must be at the service of humans. Ultimately, in order to achieve a thriving AI ecosystem that combines government funding and human-centered innovation, the best approach depends on the specific context:

- **Type of research:** Fundamental, long-term research may benefit more from government funding, while applied research and product development may be better suited for industry-led innovation.
- **Societal impact:** Projects with a strong public good component may require government support to ensure their development.
- **Market dynamics:** In areas with strong market competition, industry-led innovation may be sufficient, while government funding may be needed to address market failures or stimulate research in underinvested areas.

7. Engagement of and cooperation with the various advanced AI developers

It is not straightforward to demonstrate completely clear and unambiguous proof of deep cooperation between governments focused on AI safety and the companies actually developing advanced AI. There is a lot of activity, but it is often obscured by confidentiality, the rapid pace of change, and the inherent complexities of this emerging field.

However, there are some promising examples that suggest meaningful cooperation is happening:

1. Voluntary Commitments and Codes of Conduct:

- **AI Seoul Summit:** In May 2024, several countries signed a joint statement committing to international cooperation on AI safety research and risk management. This demonstrates a shared intent to work together, even if the specifics are still evolving.
- **Voluntary Codes of Conduct:** Many leading AI developers have signed voluntary codes of conduct related to AI safety, often in consultation with government bodies. This shows a willingness to engage with safety concerns, even if it's not legally mandated (yet).

2. AI Safety Institutes (AISI) and Joint Evaluations:

- **The UK and the U.S.:** They are actively evaluating advanced AI models with cooperation from developers. They have published individual and joint reports on their evaluations of models like OpenAI's o1, indicating some level of access and collaboration. The renamed U.S. institution will continue to '"serve as industry's primary point of contact within the U.S. Government to facilitate testing and collaborative research."
- **International Network of AISI:** This network, launched in November 2024, brings together government-backed AI safety institutes from various countries. They are working on joint testing exercises and sharing best practices, which requires cooperation from AI developers.

3. Government Funding and Partnerships:

- **Canada:** As part of the Canadian Sovereign AI Compute Strategy, the Canadian government is investing heavily in AI infrastructure and providing funding to companies like Cohere to scale up their AI compute capacity. This incentivizes companies to develop AI within Canada and potentially collaborate on safety research.
- **U.S.:** The U.S. CAISI has formed a taskforce with partners across the US government to research and test AI models for national security purposes. This suggests collaboration with developers to access and evaluate their models. The NIST AI Safety Consortium brings together more than 200 companies and researchers to jointly create test environments, data sets, guidelines, and frameworks to enable evaluation of AI models.

4. Information Sharing and Transparency:

- **International AI Safety Report 2025:** This report, written by 100 AI experts from various countries, represents a collaborative effort to synthesize knowledge about AI risks and capabilities. While it doesn't directly involve AI developers, it creates a shared understanding that can inform government policies and industry practices.
- **Increased Transparency:** There's a growing trend towards increased transparency from AI developers about their safety research and risk mitigation efforts. This openness can facilitate cooperation with governments and build public trust.

Still, there are challenges and friction points ahead that must be addressed in 2025 in order to make further progress and safeguard the Good Will among all stakeholders:

- **Confidentiality:** Much of the cooperation between governments and developers happens behind closed doors due to the sensitive nature of advanced AI technology and national security concerns. This makes it difficult to assess the full extent of the collaboration.
- **Rapid Pace of Change:** The field of AI is evolving rapidly, making it challenging for governments to keep up and establish effective cooperation mechanisms.
- **Balancing Interests:** There can be tension between governments' desire to ensure safety and developers' drive for innovation and commercial success. Finding the right balance is an ongoing challenge.

While there is no single approach that proves perfect cooperation, there is evidence that a growing trend towards collaboration between governments and AI developers on safety is becoming the best practice for everyone concerned. The level of cooperation varies, and there are still challenges to overcome, but the increasing engagement and joint efforts are a positive sign when transparency of methodologies and intent will allow governments to feel confident in industry.

It is crucial to continue building trust and transparency to foster even deeper cooperation and ensure that AI development benefits society while minimizing potential risks.

7.1 Specific actions that AI developers have put in place that exemplify their willingness and commitment to deliver a trustworthy AI

It is still early days for concrete actions that can guarantee a trustworthy AI that can move onto standardization, but there have been some promising steps that AI companies have taken to assure trustworthy AI.

1. Transparency and Explainability

- **Publishing research:** Some companies like OpenAI and DeepMind are publishing research papers explaining their AI models' architectures, training processes, and limitations. This helps external researchers understand how these models work and identify potential risks.
- **Explainable AI (XAI) tools:** Some companies are developing tools to make AI decision-making more transparent. For example, Google's Explainable AI service helps users understand why a model made a specific prediction.

2. Safety Testing and Red Teaming

- **Internal red teams:** Many companies have dedicated teams that try to find vulnerabilities or biases in their AI models. This helps identify potential problems before they cause harm.
- **External audits:** Some companies are open to external audits of their AI systems by independent organizations or researchers. This provides an additional layer of scrutiny and accountability.

3. Bias Mitigation

- **Data diversity:** Companies are increasingly aware of the importance of diverse training data to avoid bias in AI models. They are working to create more inclusive datasets that represent a wider range of demographics and perspectives.
- **Fairness metrics:** Researchers are developing metrics to measure fairness in AI systems and identify potential biases. Companies are starting to incorporate these metrics into their development processes.

4. Collaboration with Researchers and Governments

- **Sharing access to models:** Some companies are providing limited access to their AI models to researchers and government agencies for safety evaluations. This allows for independent scrutiny and helps inform policy decisions.
- **Participating in safety initiatives:** Companies are participating in initiatives like the International Network of AI Safety Institutes and contributing to the development of safety standards and best practices.
- **Voluntary code of conduct:** Frontier AI companies are engaged in the development of code of conduct with governments.

5. Responsible Deployment

- **Gradual rollout:** Companies are often taking a cautious approach to deploying new AI models, starting with limited releases or pilot programs to gather feedback and identify potential issues before wider deployment.
- **User feedback mechanisms:** Companies are implementing ways for users to provide feedback on AI systems, which can help identify and address problems quickly.

Specific Examples

- **OpenAI** conducted a pre-deployment evaluation of their o1 model with the UK and US AI Safety Institutes, demonstrating a willingness to engage with external scrutiny.
- **Cohere** is working with the Canadian government to build a new AI data center in Canada, which could facilitate collaboration on safety research.
- **Hiroshima AI Reporting Framework:** almost all the Organizations Developing Advanced AI Systems voluntarily submitted reports to share their safety initiatives and build trust.
- **Google** has published AI Principles that guide their development and use of AI, and they have established an AI Principles Review Board to provide oversight.

Challenges and Limitations

- **Commercial pressures:** Companies face pressure to innovate quickly and bring products to market, which can sometimes conflict with safety considerations.
- **Lack of standardized metrics:** There is still a lack of universally accepted standards and metrics for evaluating AI safety and trustworthiness.
- **Evolving landscape:** The field of AI and its real-world context are constantly evolving, making it challenging to stay ahead of potential risks and ensure long-term trustworthiness.

While there is still much work to be done, these actions demonstrate that AI companies are increasingly prioritizing safety and trustworthiness as a form of competitive advantage over others who do not. The willingness to engage in transparency, testing, bias mitigation, and collaboration suggests a positive shift towards responsible AI development. Continued efforts in these areas, along with the development of clear standards and regulations, will be crucial for ensuring that AI benefits society while minimizing potential harms.

8. Assuring the technical safety and trustworthiness of foundation models, Generative AI and other forms of advanced AI

To ensure the safe and correct technical functioning of a foundation model, AI developers deploy at their will a variety of (non-certified) validation techniques. Some AI institutes attempt to also validate the technical assurance of these models via proprietary methodologies that are not publicly disclosed in detail to prevent potential manipulation of the evaluation process, and are dependent on considerable governmental investment.

8.1 Transparency for AI-generated content

Making Generative AI models safe requires the content they produce to be identifiable as AI-generated. GPAI's Social Media Governance group has worked on the question of how reliable detectors for AI-generated content can be produced. They produced some influential papers and policy briefs arguing that reliable detectors are a realistic prospect if generator providers are obliged to instrument their generators to support such detection (GPAI, 2023a[19]; 2023b[20]; Knott et al., 2024a[21]). This instrumentation can be done either by having generators use watermarking schemes (see Deepmind's current watermarking scheme[22] for a recent example), or by maintaining a private log of generated content, and implementing a detector as an information retrieval system on this private log (see Krishna et al., 2023[23] for the original proposal here). A mixture of these methods is likely to be even more effective.

These arguments had some traction in the policymaking community. The 2023 GPAI report was discussed at a US Senate Judiciary Committee hearing on AI Oversight[24], where two of the authors (Yoshua Bengio and Stuart Russell) gave evidence. After discussions with policymakers in the EU Parliament and EU Commission, the work also led to additions to the EU's AI Act. In particular, Article 50.2[25] (on transparency) requires that providers of Generative AI systems 'shall ensure that the outputs of the AI system are marked in a machine-readable format and detectable as artificially generated or manipulated'. This is a significant policy impact for GPAI: we might realistically expect it will lead to more reliable detectors for AI-generated content, at least within the EU, when the relevant parts of the AI Act come into force.

This year, GPAI's social media group moved the policy discussion forward, by imagining a scenario where reliable detectors for AI-generated content are available (Knott et al., 2024b[26]; 2025[27]). Market

[19] https://gpai.ai/projects/responsible-ai/social-media-governance/Social Media Governance Project - July 2023.pdf

[20] https://oecd.ai/en/wonk/human-or-human-like-transparency-for-ai-generated-content

[21] https://link.springer.com/article/10.1007/s10676-023-09728-4?utm_source=rct_congratemailt&utm_medium=email&utm_campaign=oa_20231028&utm_content=10.1007/s10676-023-09728-4

[22] https://www.nature.com/articles/s41586-024-08025-4

[23] https://arxiv.org/abs/2303.13408

[24] https://www.judiciary.senate.gov/committee-activity/hearings/oversight-of-ai-principles-for-regulation

[25] https://artificialintelligenceact.eu/article/50/

[26] https://link.springer.com/article/10.1007/s10676-024-09795-1

[27] https://oecd.ai/en/wonk/ai-content-detection-tools

forces would likely impel certain actors to use such detectors routinely, out of pure self-interest. But other actors may be less incentivised, and new rules may be needed regarding the use of these tools. This may be the next frontier for lawmaking in relation to transparency about AI-generated content.

It is anticipated that AI will drive the advancement of co-generation. In the "From co-generated data to Generative AI: New rights and governance models" project of the GPAI Data Governance Working Group, the first step was to observe, through six case studies, the circumstances under which co-generation occurs. Following this, the project focused particularly on the EU, the United States, and Japan, analysing the legal frameworks applicable to co-generation scenarios—spanning data input and output, models, and content—primarily from the perspectives of copyright and data protection.

Further development of legal regulations on AI co-generation is essential, with key challenges including the clarification of exemption provisions, the structuring of copyright applicability to AI-generated works, and ensuring the effectiveness of data protection measures. Additionally, to serve as a reference for policymakers, the project identified seven principles under the "Principles of Copyright and Data Protection Rights in Co-Generated Input and Output of Generative AI."[28]

[28] https://gpai.ai/projects/data-governance/co-gen/

9. Creating the necessary policies and governance frameworks to guarantee the data governance of foundation models, Generative AI and other forms of advanced AI

As stated in the “Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System”, it is important to “work towards responsible information sharing and reporting of incidents among organisations developing advanced AI systems including with industry, governments, civil society, and academia.” This includes responsibly sharing information, as appropriate, including, but not limited to evaluation reports, information on security and safety risks, dangerous intended or unintended capabilities, and attempts by AI actors to circumvent safeguards across the AI lifecycle.[29]

The G7 Hiroshima AI Process (HAIP), in collaboration with the OECD, has introduced a voluntary reporting framework enabling AI service providers to disclose transparent and comparable information on their governance and risk-management practices. A preliminary analysis of this framework is presented in Ema et al. (2025)[30].

Also, like the Computer Security Incident Response Teams (CSIRTs) in cybersecurity, it is important that AI Safety Institutes around the world collaborate in sharing AI incident information and to come up with the solutions to address those issues.

The AI Incident Sharing Mechanism is a relatively new initiative aimed at improving the safety and security of AI systems by fostering collaboration and knowledge sharing about AI-related incidents. It intends to be a centralized platform where organizations and individuals can report incidents related to AI systems, such as:

- Unexpected or harmful behavior by AI models
- Security breaches or vulnerabilities in AI systems
- Bias or discrimination exhibited by AI algorithms
- Accidents or near-misses involving AI technologies

The platform allows for the sharing of anonymized incident data, protecting sensitive information while enabling the broader community to learn from these incidents. The shared data can be used for research, analysis, and the development of mitigation strategies to prevent similar incidents in the future.

The main goals for a platform such as this are to:

- **Improve collective awareness:** By sharing information about AI incidents, the mechanism aims to raise awareness of potential risks and vulnerabilities associated with AI systems.
- **Enhance AI safety:** Learning from past incidents can help developers, researchers, and policymakers improve the safety and security of AI technologies.
- **Build trust in AI:** Openly addressing AI incidents and demonstrating a commitment to safety can help build public trust in AI and its responsible development.

Several organisations have built AI Incident Sharing Mechanisms of relevance:

[29] Hiroshima Process International Guiding Principles for Organizations Developing Advanced AI System: https://www.mofa.go.jp/files/100573471.pdf

[30] Ema et al. (2025): https://www.tc.u-tokyo.ac.jp/blog/wp-content/themes/tokyocollege/publication/WP_TC-E-25-1_Ema.pdf

- **MITRE's AI Incident Sharing Initiative:** MITRE, a non-profit research organization, launched an AI Incident Sharing initiative in 2024. Their platform allows for the reporting and sharing of anonymized AI incident data. Specifically, their Center for Threat-Informed Defense, which focuses on cybersecurity and threat intelligence, is leading this effort. They have collaborated with over fifteen companies, including Microsoft, Intel, and JPMorgan Chase Bank, to develop and launch this initiative.
- **Partnership on AI's Incident Database:** The Partnership on AI, a multi-stakeholder organization focused on responsible AI, is developing an incident database to track and analyze AI harms and near-harms. This organization involves a wide range of members, including major tech companies (like Google, Meta, and Microsoft), academic institutions, and civil society organizations.
- **OECD's AI Incidents Monitor (AIM)**: The OECD is developing an AI Incident Monitor (AIM) to track and analyze real-world AI incidents and hazards. This platform is being informed by the work of the OECD.AI expert group on AI incidents, which is working on defining AI incidents and associated terminology. The AIM aims to provide a "reality check" to ensure that the reporting framework and definitions work in practice. It uses machine learning models to classify incidents and hazards based on various factors like severity, industry, and related AI principles.
- Other initiatives: Several other organizations and research groups are working on AI incident sharing mechanisms or databases, including the Center for Security and Emerging Technology (CSET) at Georgetown University, who advocates for a hybrid AI incident reporting framework that includes mandatory, voluntary, and citizen reporting mechanisms, and the *AI Now Institute* at New York University is also involved in research and advocacy related to AI incidents and accountability.

Creating centralized platforms for incident reporting is a challenging endeavour specially when it comes to data sensitivity, where balancing the need for data sharing with the protection of sensitive information is not straightforward. Furthermore, the success of such platforms can only be guaranteed if standardized reporting formats and definitions for AI incidents are developed for effective data collection and analysis. Likewise, creating incentives for organizations and individuals to report AI incidents is essential for the success of these mechanisms, especially when the aim is to report them accurately, and with trust and transparency among the participants.

The challenge with incident reporting platforms is they are not AI system specific. It is potentially important to measure AI subsystem level failures as a standard protocol to measure and detect failures within and across AI systems at scale. Many critical engineering systems such as aerospace, nuclear power plants, military and defence equipment have well established standards and measurement for failure reporting specific to the system in question. It is potentially possible to extend the system status pages (that are already reported by OpenAI, Anthrophic, GCP, AWS etc..) to build integrated data repositories that are system specific to inform prognostics and health monitoring (PHM) to measure, detect and proactively prevent the varied modalities of AI failures. In similar vein, tried and tested methodologies in Accelerated Life Testing (ALT), reliability, resilience and human factors engineering methods can be applied to AI systems at a component level to bring visibility to modes of failure, root causes to help inform community-wide proactive approaches to fix critical safety and reliability for the whole AI system.[31]

[31] https://arxiv.org/pdf/2411.08981

The aim of any incident reporting should be to offer enough granularity of information that allows for the identification of the nature of the failure (human, data, model or code and software, computing environment failure).

10. Conclusions and recommendations

The activities of the AISIs can only succeed via the close collaboration of all stakeholders in the AI ecosystem: national government agencies, such as national security agencies, ministries of defence, commerce, science and technology; centres for ethics, data governance, ombudsman, government accountability office, watchdogs, inspector generals and similar bodies that seek fairness in society; academic institutions researching artificial intelligence; and the companies developing and commercialising advanced AI.

The recommendations offered in this paper have been based on feedback derived from experts engaged in academic research, development of AI policy, and industry commercialising AI.

There is much hesitation by governments when it comes to Generative AI due to the pressure of the companies developing it, the confusing messages in the marketplace exchanged between academics that reject some of the developments and funding and investment pressures The authors of this paper propose recommendations that are strategic and predictive of market evolutions and their potential consequences, while clarifying often-misunderstood concepts.

Recommendation 1: **Bridging AI safety and AI fairness**

Amplification of known risk factors such as bias, black box nature, privacy, cybersecurity and the emerging risks from Generative AI (i.e. disinformation, safety, malicious use) will require governments to bridge the AI safety and AI fairness agendas. AI safety agenda cannot succeed to the detriment, or in the absence of ignoring impact on human rights. AI systems should be unbiased and trustworthy. Risk and capability indicators must take into account the level and scale of risks to individuals and society too.

Recommendation 2: **Strengthening the International Cooperation Above Geopolitical Shifts**

The International Network of AI Safety Institutes must have a robust coordination and cooperation mechanism. Such a mechanism must advance the science of safety, depth of evaluations, knowledge-sharing, and incident response. The Network can strengthen the institutional capacity of each member, while making the Network itself more resilient to geopolitical changes and/or corporate actions. A network can be stronger than the sum of individual members, and create new synergies for interoperability of AI governance frameworks.

Recommendation 3: AI Safety is the pillar upon which to build the future society

We agree with the conclusions of the recently published International AI Safety Report, in that “nothing about the future of general-purpose AI is inevitable” and who benefits from AI and what risks are acceptable will depend on "the choices that societies and governments make today and in the future to shape the development of general-purpose AI.”[32]

Recommendation 4: Human Creativity should be protected as a human right

AISIs should also delve into assessing the future of human creativity in a world where potentially Generative AI is widely adopted as a common denominator for simple tasks, but still aiming to overtake human creative artistry, as AI developers continue to train AI systems to produce works of art rather than industrial designs. The new scope towards dealing with the future of creativity cannot

[32] “International AI Safety Report” (DSIT 2025/001, 2025); https://www.gov.uk/government/publications/international-ai-safety-report-2025

be solely understood via the single lens of ethics, but from the perspective of its effect in human neurodivergence, and its evolutionary path to superior cognition. Neuroscience has demonstrated the benefits of teaching and encouraging humans towards creative endeavours as a form of mental and emotional health, and the untapped creative capabilities of every human, which are equal or perhaps more important to develop than other areas in knowledge and physical education.

Recommendation 5: Preventing freedom of speech becoming content radicalization
Democratic governments may be forced to conduct stricter enforcement on social media as the social media companies have ceased to do it within their internal operating processes. It is of utmost importance that freedom of speech does not become content radicalization, which is a major threat to society and a widespread issue in social media, especially if this content is false and undermines societal values such as inclusivity or creates racial or religious profiling that may harm people.

Recommendation 6: Increasing the monitoring of agentic AI development
Unless the development of agentic AI is addressed head on by governments as of immediate effect, versions of this AI systems will be commercially deployed until developers are stopped by regulators, which explains why current companies with AGI developments have re-capitalised (Anthropic raised a further US$ 1 billion on 22 January 2025 from Google), and accelerated the development of AGI. Governments must realise sooner rather than later of the unstable environments in which agentic AI is being further automated and challenge companies to stop their developments.

Recommendation 7: Incident reporting should adopt system specific approaches to AI
Industry and service operations are increasingly dependent on AI systems, so it is imperative to ensure their reliability and safety. Monitoring frameworks must integrate reliability and resilience principles into AI systems similar to those that exist in engineering systems. Not only traditional metrics, such as failure rate and Mean Time Between Failures (MTBF), must be accounted for but also human reliability because real-world AI systems still encompass human intervention.

Recommendation 8: Aiming for Standards
If AI is to become a solid foundation of innovation, it must offer reliability in its functionalities. Every aspect that makes AI a system of future architectures, processes, features and tools used by humans, must align with what other transformative innovations did to become digital infrastructures of progress. This may be challenging, but not impossible.

11. Addendum

As Generative AI models continue to be trained, the challenges have also grown exponentially and their repercussions are now widespread across all areas of model safety and trustworthiness evaluation. Compounding these issues is the lack of active governmental funding of open-source, public AI models and datasets—leaving most advances and associated safeguards to the private sector—which further complicates efforts to ensure that these technologies can be trusted to serve the broader public interest. To put it in simple terms: we find ourselves in a step-back position rather than out of the storm. As the models grow in size, the obstacles for their training have become not merely bottlenecks but directly influence the safety and trustworthiness of the resulting AI-based systems. In parallel, the emergence of agentic technologies—AI-based systems that act with a degree of autonomy, setting their own goals and executing complex tasks without direct, granular human oversight—has introduced unprecedented challenges and amplified existing concerns about safety, accountability, and ethical alignment. As these systems grow in complexity and autonomy, the risks expand not linearly, but exponentially: the very features that make agentic systems powerful also render it unpredictable, opaque, and—if unchecked—potentially disruptive across society.

These two contexts of development will have an unavoidable effect on any policy-making or standardisation efforts, because models will continue to evolve faster than governments can produce any solution based on regulation, or principles. The core nature of Generative AI models is revealing unexpected properties that increase the challenge of training them under the safety and trustworthiness values, principles and objectives in both short and medium timeframes. The technical advances that drive the capabilities of agentic systems forward put the surrounding frameworks for governing, understanding, and securing these systems at a pace and a strain that leaves them lagging dangerously behind.

Understanding the new dimensions of risk

The magnitude of the fast-paced evolutionary path that both Generative AI and Agentic technology developments have undergone in recent months is worth detailing.

On the one hand, Generative AI presents clear challenges for three specific areas of validation:

1. Data-Centric Obstacles

Model training is suffering from a scarcity of high-calibre data, forcing AI-based development companies to cut corners in acquiring quality datasets. In doing so, they have (un)knowingly trained models with substandard data which now present biased, inaccurate, or irrelevant model outputs. In addition, regulatory restrictions in data acquisition such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA) have forced the training of models towards environments where the data has created narrow scopes, and outdated information, forcing models towards hallucinations or the creation of synthetic data in order to complement its cognition training. These two well-documented drawbacks have recently expanded to additional contexts of data challenges:

A. Data annotation still demands extensive human effort (higher risk of manual errors) or highly specialised tools not available to all AI-based development companies;

B. The danger of “*data hungry*” models challenge companies to bypass social responsibilities (what may be legal may not be socially responsible to use) due to data restrictions;

C. The “data paradox” is now challenged by the weakness of the argument that “the more data an algorithm is trained on, the more accurate its predictions become” when pursuing volume of data also dangerously increases and amplifies a multitude of ethical and legal risks (heightened privacy and security concerns, complexities around data consent, inclusion of low quality or inaccurate data, increased potential for intellectual property infringement, and the perpetuation of biases embedded within the data) thus introducing a new layer of complexity for data governance and risk management;

While neutral and accurate outcomes is the goal, volume over quality approach might undermine these efforts;

2. Model-Centric Complexities

Current model architectures, by their nature, can generate incorrect, misleading, or entirely false outputs. LLMs in particular continue to fail because of their probabilistic nature in “assuming the world” rather than “knowing the world” limiting their absolute utility (and challenging their standardisation as reliable tools for industries);

In addition to the hallucinations, there are a new set of model malfunctions worth highlighting (model collapse; recursive risk; knowledge decline; catastrophic interference; the black-box wall). The “opacity” of model inference is very risky to assume in highly regulated sectors such as healthcare or finance, increasing public mistrust in AI-based systems;

3. Operational & Development Obstacles

Inherent training instability of Generative Adversarial Networks (GANs) lead to “model collapse” where outputs become repetitive and fail to capture the full diversity of the data distribution, creating a narrower funnel of knowledge. Mitigating this fault requires human monitoring and models fail to create within their learning knowledge nuances. This reveals that the instability of the models is not purely algorithmic, but reflects the increased difficulty of aligning human real-world perception and machine interpretation of the same reality. Unsurprisingly, there is an enormous scarcity of qualified professionals that can become reliable model trainers, which become more and more complex and in need of more skilled and specific training to continue their learning curves;

On the other hand, training agentic systems to handle the real-world means preparing it for an almost infinite array of unpredictable scenarios:

A wider scope of the real world and the tasks to fulfil

Unlike earlier forms of automation confined to narrow, well-specified domains, agentic systems must navigate ever-changing, ambiguous circumstances, from urban traffic to sensitive healthcare decisions. Nevertheless, no reasonable training set can fully encapsulate the nuance of lived human reality, making incomplete coverage and edge-case failures nearly inevitable. The ethical dimension is even more concerning. Autonomy in agentic systems means these technologies will increasingly make consequential/critical decisions without the practical possibility of real-time human intervention. Many of these choices touch on issues where reasonable people can disagree, and where context is everything. Embedding "human values" is not a solved problem—values are diverse, situational, and contested—and failure to align AI-based technologies' behaviour with ethical expectations poses risks of unfairness, harm, or unintended discrimination, especially for vulnerable groups. One must consider that bias present in training data or reward design is likely to be magnified by autonomous optimisation and scale quickly if not diligently governed.

Accountability grows muddier as agentic technologies proliferate.
When autonomous agents make decisions with significant real-world impact—say, in medical triage, legal reasoning, or industrial control—responsibility for failures becomes diffuse. Are the engineers who designed the system at fault, or the data providers, or the business deploying the system, or the AI itself? The opacity of complex agentic architectures, often described as black boxes, compounds the problem by making it nearly impossible to trace decision logic or contest adverse outcomes. In high-stakes domains, this lack of transparency erodes trust and raises serious legal, societal, and regulatory questions.

Technical and operational risks that compound these foundational challenges.
Agentic technologies often rely on self-supervised or reinforcement learning with reward systems that are difficult to specify comprehensively. Poorly specified incentives can cause agentic systems to "exploit loopholes," producing superficially effective but ultimately hazardous behaviours. Systems composed of multiple autonomous agentic elements introduce numerous failure points; miscoordination or single-point errors can cascade catastrophically through interconnected environments. At the organisational level, adoption is further hampered by skill and literacy gaps, overwhelmed governance mechanisms, and a lack of robust oversight—pitfalls that leave even well-resourced enterprises exposed.

Asynchronous pace between innovation and regulation
Intensifying these concerns is the reality that regulation, transparency, and safety protocols are not keeping pace with deployment. As agentic technologies outstrip traditional human control frameworks, even organisations seeking responsible governance struggle to build adequate monitoring, fail-safes, and bias assessments into their rapidly evolving structures. As a result, we face a future where robust autonomous systems operate at scale without proven strategies for keeping them trustworthy, robust, fair, and accountable.

In sum, for all the promise of agentic technologies, the current trajectory demands urgent caution. The problems presented are not merely bottlenecks—they are fundamental threats to the safety, legitimacy, and trustworthiness of increasingly autonomous AI systems. Without a coordinated focus on holistic governance, ethical alignment, transparency, and robust technical oversight, the spread of agentic AI risks amplifying harm and confusion rather than delivering the widespread, beneficial transformation its advocates promise.

Progressing SAFE Project to respond to these challenges
How we propose to continue monitoring and evaluating these developments is to focus on :
a) What safety and trustworthiness risks originate in training to continue fine-tuning the multidimensional framework of what trustworthiness means where it comes to fairness, robustness, explainability, transparency, privacy, and accountability;
b) What strategies and methodologies for mitigating training challenges and enhancing trustworthiness align themselves with the objectives of policy makers and industry standards to prevent the propagation and amplification of bias; the substantial risk of massive misinformation creation and distribution; increase of data and cyber security vulnerabilities; unlawful use of

intellectual property and copyrighted materials; and rising broader societal and ethical implications, particularly enablement of malicious actors;
c) Propose potential avenues towards a responsible future for both Generative AI and Agentic technologies that is feasible and sensible to both innovation and equitable progress within legal frameworks – from active data curation strategies designed to detect not only technical and statistical bias but also culturally embedded biases, to improved Privacy-Enhancing Technologies (PETs) and scalable data lineage and provenance tracking that industry can adopt and auditors can effectively monitor; all accompanied by methodologies that actively synchronize the pace of innovation with the speed of policy making to ensure continued vigilance against emerging forms of bias and inequity.

The challenges described highlight that the collaborative efforts between academia, industry, and policy makers are crucial to navigate the full comprehension and extent of impact of Generative AI and advanced AI-based systems. Transparency and collaboration between stakeholders (public, academia, governments, and the private sector) allow for jointly determining what we shall trust in the future as an AI that guarantees wellbeing and the stability of our society.

In addition, we cannot discard the reality that some of the challenges will be known to the public, but many others may not be shared by AI-based development companies in full or in time for policy makers to comprehend the entirety of their effect in the world.

Keeping this project as an ongoing research activity within the GPAI activities outside of the formal scope of projects, but under the GPAI's mission of engaging with and monitoring the development of AI-based systems beyond policy, will be crucial to ensure an open channel between AI-based systems innovation and cutting-edge development and their alignment with the values, principles and objectives that drive service-led industry standards and functional policy making.