GOUVERNEMENT
Liberté
Égalité
Fraternité

PRESS RELEASE

Paris, 12/02/2025

AI Action Summit: Ensuring the development of trusted, safe and secure AI to benefit of all

Artificial intelligence (AI) is profoundly transforming our society, opening up unprecedented opportunities in many fields. For this technological revolution to benefit everyone, it is crucial to ensure its responsible and ethical development, with trust at its heart. AI raises major concerns regarding safety and security, as clearly demonstrated by the summits in Bletchley Park (United Kingdom - November 2023) and Seoul (South Korea - May 2024). Whether in anticipating extreme risks or addressing those already visible, a resolutely ambitious approach to building trust in AI is essential on an international scale.

Technological advances in AI also offer exceptional possibilities in the field of security. With this in mind, the AI Action Summit is committed to promoting safe and secure AI, particularly by providing the necessary tools to mitigate these risks.

For AI to fulfill its promises, a collaborative approach is essential. The AI Action Summit calls on public, private, and academic stakeholders to work together to build a trusted AI ecosystem.

This global approach is based on three pillars: **science, solutions, and standards**. With a robust international scientific consensus on AI, the time has come to develop technical solutions that are open and accessible to all, while creating common international standards recognized by the entire ecosystem. This will help prevent fragmentation and encourage convergence at all levels.

As part of the AI Action Summit, the Ministry of Europe and Foreign Affairs and the General Secretariat for Defence and National Security have supported the action of the thematic envoy, Guillaume Poupard, to federate an ecosystem of stakeholders, both nationally and internationally, mobilised to strengthen the safety and security of AI.

AI ACTION SUMMIT

ASSESSING AND RESPONDING TO THE SAFETY AND SECURITY CHALLENGES OF ARTIFICIAL INTELLIGENCE

AI Safety Report.

Following the Bletchley Park and Seoul Summits, work on the risks of advanced AI models has continued under the leadership of Yoshua Bengio, supported by a panel of renowned international experts. The conclusions of this report, officially presented at the Summit, paved the way for in-depth discussions between global experts during the scientific sessions and throughout the event.

The report aims to describe how general-purpose AI models function, identify the risks they may pose, and explore ways to mitigate them. It will be continually updated to reflect the latest advancements, under British secretariat oversight.

> International AI Safety Report 2025

Launch of INESIA: The French Institute for AI Evaluation and Security.

On **January 31**, **Clara Chappaz**, the French Minister for AI and the Digital Economy, announced the creation of a **national institute dedicated to AI evaluation and security**: **INESIA**.

Led by the **General Secretariat for Defence and National Security (SGDSN)**, on behalf of the **Prime Minister**, and the **Directorate General for Enterprise (DGE)**, this institute will unite top national players to address AI challenges.

Through research, evaluation, and international collaboration, INESIA aims to promote **human-centric and responsible AI**. This also involves developing **robust technical tools** to guarantee security and mitigate AI-related risks.

INESIA joins the growing network of **'AI Safety Institutes'** launched at the **Bletchley Park Summit**, now active in about **ten countries** and forming a **global network** in **San Francisco since November 2024**. The first joint experimental findings were unveiled at the Summit.

As part of the network continued effort to advance the science of AI model evaluations and work towards building common best practices for testing advanced AI systems, Singapore, Japan, and the United Kingdom led the Network's latest joint testing exercise aimed at improving the efficacy of model evaluations across different languages.

This novel international collaboration is key to ensuring that model evaluations – especially in public safety and national security domains such as cybersecurity – are robust, accurate, and account for the nuances of global languages. France provided the dataset that helps to conduct one of the two cybersecurity evaluation.

> Joint testing

Launch of the French Language Model Leaderboard.

To enhance transparency, rigor, and independence in AI evaluation, **the French national AI coordination**, in collaboration with **the Ministry of Education, Inria, LNE, and GENCI**, has partnered with **Hugging Face** to create a **French-language model leaderboard**.

This **trusted repository** enables structured comparisons of AI models based on objective and transparent criteria. Its objective is clear: **to improve the French-language performance of the best language models, regardless of where they were developed.** This will ensure that all French-speaking companies and institutions have access to the most advanced technologies, optimised for their working language and specific uses.

> French AI Model Leaderboard

AI REGARDING FOREIGN DIGITAL INTERFERENCE, CYBERSECURITY AND DATA PROTECTION

Combating Information Manipulation.

The rise of AI, combined with the algorithmic mechanisms of social networks, raises concerns about its impact on the integrity of information. This combination of technologies carries the risk of distorting citizens' perception of reality, potentially undermining democracy, which is built on trust.

As part of the Summit, a coalition of French and international stakeholders from the public, private, institutional, and academic sectors has pooled its expertise to explore both the challenges and opportunities of AI in safeguarding information integrity and countering information manipulation.

To foster dialogue on these critical issues, **VIGINUM and the OECD**, with the support of the **Ministry of Culture**, hosted an official side event on **February 11**.

To support this initiative, VIGINUM published a report on the **challenges and opportunities of AI in combating information manipulation**. This report, incorporating insights from international partners, is built around a key idea: while AI technologies may heighten the threat of foreign digital interference, they also provide powerful tools to enhance our defenses.

> VIGINUM report

To translate these possibilities into action, **VIGINUM** released an open-source tool, **'D3lta'**, designed to detect large-scale text manipulation tactics using AI.

> D3lta on GitHub

A second tool, developed by **PEReN** (the Center of Expertise for Digital Platform Regulation) in collaboration with **VIGINUM**, uses a **meta-detector** to evaluate the performance of various AI-generated content detection systems. By aggregating different detection methods, this tool aims to identify artificial content that might otherwise go unnoticed.

> AI Summit Tool - PEReN

Additionally, a study on the practical applications of AI tools in journalism—conducted by journalists for their peers— was published at the Summit.

The Summit also saw the **creation of a global fact-checking network**. Furthermore, a **joint media declaration** outlined specific measures to uphold the integrity and control of trustworthy and professional information, in defense of democracy.

Renforcer la cyber-résilience.

The large-scale deployment of AI technologies presents major cybersecurity challenges on three fronts:

- **Securing AI systems** by developing evaluation frameworks and supporting stakeholders in their AI projects,
- **Enhancing cybersecurity tools** with AI-driven capabilities, and
- **Raising awareness of AI's potential malicious uses**.

In collaboration with national and international partners, **ANSSI** has been actively identifying and analyzing AI-related cybersecurity risks to help secure AI systems and promote their responsible deployment.

An **international risk analysis** was presented on **February 7** during the **scientific sessions organized by the Institut Polytechnique de Paris**. This research highlights the urgent need to adapt to the cybersecurity challenges posed by AI.

Based on this analysis, a **crisis simulation exercise** has been held on **February 11**, bringing together over **200 AI and cybersecurity experts**. This exercise facilitated dialogue and tested operational scenarios to identify defense and governance measures for securing AI-integrated information systems.

A **post-exercise report** summarizing key lessons learned will be published after the Summit, contributing to a broader culture of AI cybersecurity crisis management.

> AI through a cyber risk-based approach

Protecting privacy.

The development of AI technologies raises key issues in terms of privacy. Indeed, many systems use personal data both for their development and during their use. Data protection rules provide a clear and protective governance framework that fosters trust. This is particularly important in the current context of development and deployment of AI, where data processing is becoming extremely complex, mobilising a highly fragmented value chain, operating on a very large scale and entailing major risks of opacity for individuals.

A number of events took place at the Summit to shed light on these issues from different angles and to provide answers, in particular with regard to:

- **Data protection**, particularly with regard to the issues surrounding voice and image processing in the age of AI;
- **Governance**, by considering privacy issues at a global level in a globalised technological environment, clarifying requirements for stakeholders, facilitating exchanges between regulators and providing security for individuals;
- The **practical implementation of European regulation**, at a time when many normative texts are coming into force: General Data Protection Regulation (GDPR), AI Act, Digital Services Regulation (DSR), etc.

TRUST IN AI GOVERNANCE.

To balance **innovation with ethical AI development**, the **European Union adopted the AI Act in 2024**, the world's first **binding** and **comprehensive regulatory framework**. With key provisions set to take effect in 2025, the European Commission is working closely with industry players to develop a **Code of Practice** for proper implementation.

At the international level, the **Hiroshima Process Code of Conduct**, negotiated by the **G7 since 2023**, has been officially launched alongside the Summit. This initiative encourages AI developers worldwide to uphold essential ethical and safety principles.