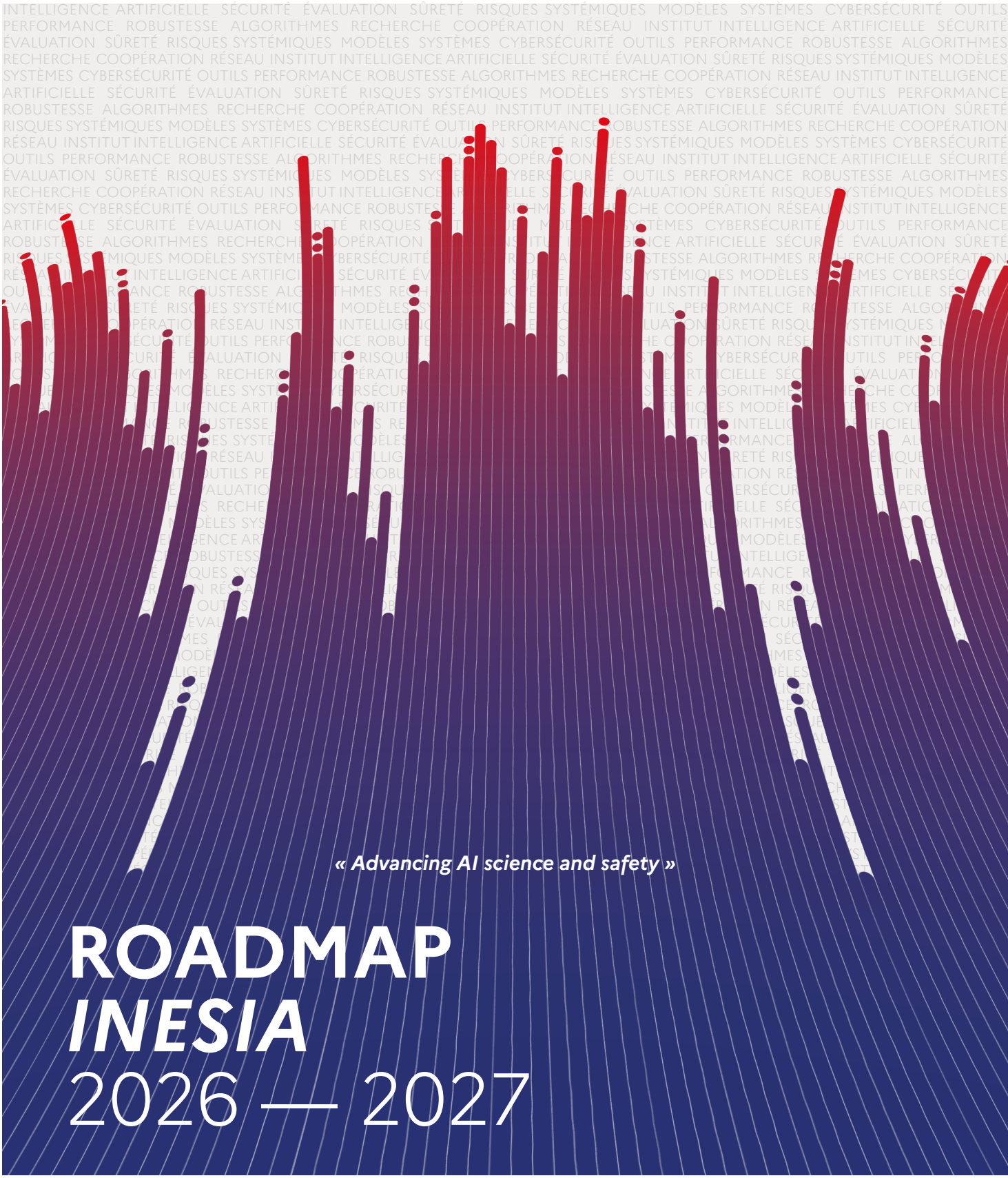
GOUVERNEMENT

Liberté
Égalité
Fraternité



« Advancing AI science and safety »

ROADMAP *INESIA* 2026 — 2027

Courtesy translation. The French document is the official version.

EXECUTIVE SUMMARY

AI systems are reaching new levels of complexity. Equipped with multimodal reasoning capabilities, able to interact with digital and physical environments, and even adapt continuously, the most advanced AI models are becoming ubiquitous in all areas of society. Already integrated into public services, research, education, and professional tools, they are set to play a role in decisions with major social, economic, and security implications.

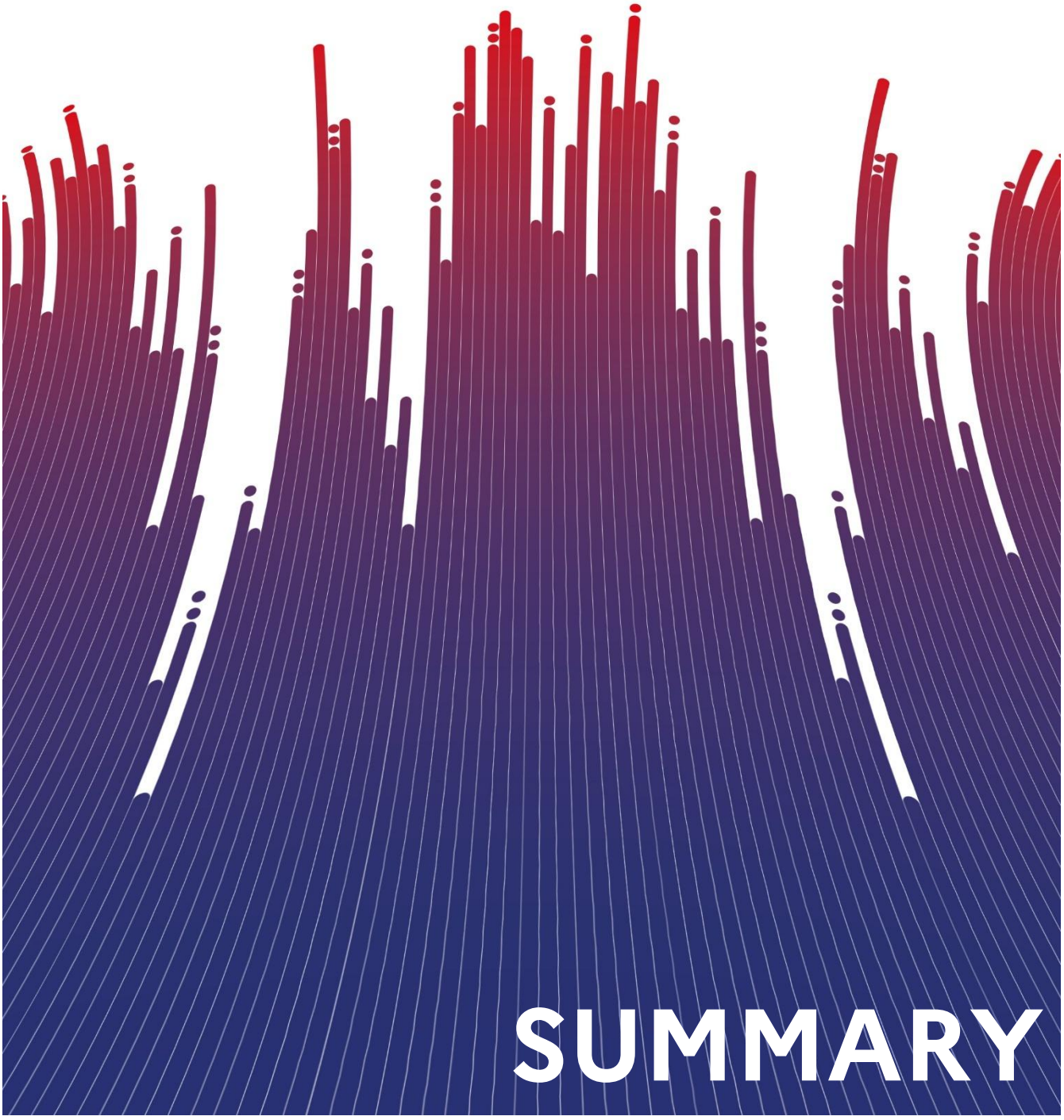
These developments are giving rise to a series of unprecedented challenges. The ability of AI to generate content that is indistinguishable from reality, to influence human behaviour, or to orchestrate actions in complex systems accentuates its systemic potential. Designed for closed, static, and single-function systems, traditional evaluation methods appear outdated in the face of this.

The evaluation of AI is therefore becoming a strategic issue in its own. It is not limited to measuring performance: it must enable the detection of unexpected behaviour, the quantification of uncertainties, and the evaluation of robustness, alignment, resilience, and the risks of misuse. It draws on a demanding interdisciplinary field, combining data science, behavioural modelling, IT security, cognitive ergonomics, and the philosophy of technology.

In this changing landscape, at the intersection of technological dynamics and sovereignty requirements, the National Institute for the Evaluation and Security of Artificial Intelligence (INESIA) was created. Launched in 2025, it brings together ANSSI, Inria, LNE, and PEReN. Its mission is to structure a public capacity for evaluating advanced AI that is rigorous, independent, and, where necessary, interoperable with European and international initiatives.

INESIA operates at the intersection of three dimensions: supporting regulation in support of the European regulation on artificial intelligence (AI Act)'s implementation, controlling the systemic risks of AI, and supporting the evaluation of the performance and reliability of models and systems. Its work is conducted methodically: understanding emerging systems, testing their behaviour, structuring evaluation tools, influencing standards, and guiding collective choices.

By pursuing this roadmap, INESIA intends to lay the foundations for a robust, open, and sovereign AI evaluation capability that can inform public decisions, strengthen collective resilience to risks, and ensure that technological innovation takes place within a framework of trust.



SUMMARY

INTRODUCTION

FINDING

The pace of progress in AI shows no sign of slowing down. The development of models capable of reasoning, manipulating multimodal data, interacting with hybrid environments, and soon self-improvement, poses new challenges. These challenges raise questions about performance and interpretability, as well as safety, robustness in the face of the unexpected, and alignment with human values and intentions.

Evaluation methods designed for closed, single-function, or static systems are showing their limitations. New benchmarks need to be developed to assess their robustness, resilience, contextual reliability, and the risks posed by misuse. As **AI evaluation becomes a scientific field in its own right,** it is essential that France has the technical expertise to identify, measure, and contain risks with high systemic potential, including those related to information manipulation, cybersecurity, or the destabilization of collective processes.

RESPONSE

INESIA aims to contribute to this "science of evaluation." Created in 2025, INESIA's mission is to build a sovereign and public capacity for evaluating advanced AI by structuring a rigorous knowledge base, bringing together the relevant stakeholders, and facilitating exchanges between academia, government, industry, and civil society. It aims to structure a sovereign AI evaluation capacity at the intersection of scientific excellence, technical innovation, and operational impact requirements.

INESIA provides technical expertise to government departments and promotes responsible innovation. It acts as a national anchor point in trusted international networks and, as such, represents France within the AI Safety Institutes network.

INESIA's work is based on four fundamental principles: **developing a scientific and strategic approach, structuring a sovereign evaluation capacity, coordinating public expertise and pooling resources, and being visible and active in international networks**. Operationally, this approach involves gaining a better understanding of the systems to be evaluated (through new metrics and new tests), testing further and faster (through robust technical tools and platforms), disseminating this knowledge both within the scientific community and in the regulatory ecosystem, and guiding the evaluation practices of the most powerful future models. This ambition is organized around three strategic areas and one cross-cutting theme.

POLE 1 – SUPPORT FOR REGULATION

TO CONTRIBUTE TO THE BALANCED AND EFFECTIVE IMPLEMENTATION OF THE AI REGULATORY FRAMEWORK, WHICH PROTECTS CITIZENS AND SUPPORTS INNOVATION.

INESIA will contribute to providing the authorities with the evaluation tools they need for their activities (**Project No. 1**). Regular progress reports will enable monitoring of the regulatory work linked with the implementation of the AI Act (**Project No. 2**). In addition, work already underway in the field of synthetic content detection will be continued (**Project No. 3**). Finally, evaluation methods adapted to the cybersecurity of AI systems and cybersecurity products incorporating AI will be developed (**Project No. 4**).

POLE 2 – SYSTEMIC RISKS

TO OBJECTIVELY ASSESS THE SYSTEMIC RISKS THAT AI COULD POSE, INFORM PUBLIC DECISION-MAKING AND AFFIRM FRANCE'S POSITION IN AI SECURITY.

Research activities will be conducted to better understand systemic risks, characterise them, control them and inform public decision-making (**Project No. 5**). Further downstream, work will be carried out to refine the evaluation of agentic systems (**Project No. 6**). Finally, the continuation of activities conducted within the framework of the international AISI network will affirm France's position in efforts to promote AI safety (**Project No. 7**).

POLE 3 – PERFORMANCE AND RELIABILITY

TO STIMULATE CREATIVITY IN THE ECOSYSTEM, USE EMULATION AS A LEVER TO PROMOTE THE EMERGENCE OF MORE EFFICIENT AND RELIABLE SOLUTIONS.

Challenges will be conducted to strengthen competition between stakeholders and encourage the development of creative solutions (**Project No. 8**).

CROSS-CUTTING AXIS

TO CREATE STRUCTURING COMMONS FOR INESIA'S WORK, DEVELOP SYNERGIES BETWEEN THE PARTIES AND THEIR WORK, AND BRING TOGETHER AN ECOSYSTEM.

Pooling best practices and knowledge will enable us to capitalise on existing resources and strengthen our knowledge (**Project No. 9**). It will be necessary to set up the technical architecture that will enable the evaluation of models and systems (**Project No. 10**). Finally, the organisation of scientific outreach activities will help to bring together the AI evaluation and security ecosystem (**Project No. 11**).

In line with this vision, this roadmap sets out INESIA's priorities in concrete actions scheduled over time. It has been developed in conjunction with the needs identified by the relevant authorities, European guidelines, ongoing international initiatives and leading scientific research.

SUPPORT FOR REGULATION PILLAR

Contribute to the effective implementation of the AI regulatory framework, for a balanced application that protects citizens and supports innovation.

INESIA aims to provide technical expertise to AI regulatory authorities. Given the increasing complexity of systems, the diversity of their uses and regulatory requirements, it is essential to design systematic, replicable testing protocols that are tailored to the specific needs of public authorities.

Project No. 1 aims to contribute to **the provision of evaluation tools for market surveillance authorities** that will monitor high-risk AI systems as part of the implementation of the AI Act. With a view to efficiency, existing tools will be listed and compiled, and research may be undertaken to develop any missing tools.

Acculturating INESIA members to regulatory issues, particularly in the field of evaluation, will enable discussions on the ongoing deliberations within the various standardisation bodies: ISO-IEC, CEN-CENELEC, etc. (Project No. 2).

Project No. 3 will continue the work already underway to accurately assess the performance of **synthetic content detectors in real-world conditions**, a key issue in ensuring information security. As an extension of this, the capabilities for detecting AI-generated content will be strengthened by expanding and improving the reliability of an open-source reference library.

Finally, the SEPIA project (Security of Products Integrating AI, Project No. 4) will **develop certification methods tailored to the cybersecurity of AI systems and cybersecurity products incorporating AI**, intended for the certification ecosystem, and designed to provide a harmonised response to the various regulatory frameworks concerned (European regulation on AI, European regulation on cyber resilience, European regulation on cybersecurity).

PROJECT N°1.

MAKING EVALUATION TOOLS AVAILABLE TO NATIONAL REGULATORS

OBJECTIVE

Provide French market surveillance authorities with evaluation tools under the AI Act, by mapping existing relevant tools on the one hand, and by developing missing tools in line with their needs on the other.

CONTEXT

As part of the implementation of the AI Act, several competent authorities (known as market surveillance authorities) will be responsible for monitoring high-risk AI systems placed on the French market. They will verify the compliance of these products with various technical and administrative requirements.
The competent authorities in France would benefit from a list of evaluation tools, for the sake of efficiency and harmonisation of practices, and a framework for developing any missing methods.

METHODOLOGY

As part of the planned technical support framework for the national implementation of the AI Act, PEReN plans to gather information on the needs of market surveillance authorities. Contact with European counterparts may also be proposed. The lessons learned from these exchanges will be incorporated into the project. At the same time, the project members will conduct a survey of existing tools and resources, which will lead to the identification of priority areas for research and development, based on the gaps identified in relation to needs. Priority evaluation tools and methodologies that are lacking may be the subject of research or development activities, in line with the harmonised standards planned to support the AI Act.
The project may be conducted in conjunction with Project No. 10, as necessary.

DELIVERABLES

- Summary of regulators' needs and identification of areas requiring INESIA's assistance through research and development work.

STAKEHOLDERS

ANSSI, Inria, LNE et PEReN.

PROJECT N°2.

PROMOTING DISCUSSION ON ISSUES RELATED TO AI STANDARDISATION

OBJECTIVE

Promote exchanges between INESIA members on standardisation issues, particularly in relation to the evaluation and security of general-purpose AI models, as well as responding to the demand for standardisation in support of the AI Act.

CONTEXT

The European Commission has launched a standardisation request in support of the AI Regulation's compliance requirements for high-risk AI systems, which is still ongoing. In addition, it will also submit a standardisation request for general-purpose AI models, including issues such as their safety, evaluation methods and environmental footprint.
Other frameworks such as the OECD, the AISI network and the Hiroshima process are working to define common frameworks for testing the robustness, reliability, transparency and cybersecurity of AI systems.

METHODOLOGY

Heavily involved in monitoring the implementation of the AI Act and focused on monitoring standardisation work, the DGE will be able to organise progress reports on harmonised standards to ensure information sharing among INESIA members. It will also be able to alert the various stakeholders to standardisation issues of interest outside this cycle of meetings, as necessary.
Depending on opportunities, this work may include other public stakeholders active in standardisation bodies on a broader basis.

DELIVRABLES

- Organisation of quarterly meetings to disseminate information on the work in progress and inform stakeholders of regulatory developments.

STAKEHOLDERS

DGE, ANSSI, LNE et PEReN.

PROJECT N°3.

EVALUATING THE DETECTION OF ARTIFICIAL CONTENT IN REAL-WORLD CONDITIONS

OBJECTIVE

Strengthen the evaluation capabilities of detectors of artificially generated content (text, image, audio, video) by developing a regularly updated reference library that allows the performance of detection tools to be compared under conditions close to those encountered in operational use case.

CONTEXT

Advances in generative AI are leading to the rapid spread of synthetic content that will soon be indistinguishable from authentic content, posing a risk to trust in information: disinformation, interference, mass manipulation. It is strategic for public authorities to have the capacity to detect such content.
However, most existing work is based on laboratory tests that are not representative of the diversity of content actually disseminated on the Internet. In 2024, PEReN and VIGINUM began developing an open-source library dedicated to evaluating detectors of generated text and images.
Released as open source at the Paris AI Action Summit, this foundation is a first step forward. The task now is to maintain it by actively monitoring the state-of-the-art to ensure its scientific and operational relevance, and to extend it to other critical modalities such as audio and video.

METHODOLOGY

The project will comprise three main components:
- review of the state of the art: inventory and critical analysis of methods for detecting artificially generated audio and video content;
- extension of the detection library: integration of new modalities (audio, video) and addition of relevant detectors from the state of the art, as an extension of the existing code co-developed by PEReN and VIGINUM;
- if appropriate, exploration of ensemble approaches combining several tools for performance purposes.

If necessary, non-public evaluation sets may be created.

DELIVRABLES

- Updated state of the art in generated content detection;
- Enriched code base;
- Evaluation results on non-public datasets for dedicated benchmarking purposes, i.e. under real-world conditions.

STAKEHOLDERS

ANSSI and PEReN. VIGINUM as external partner.

PROJECT N°4.

METHODS FOR EVALUATING THE CYBERSECURITY OF AI SYSTEMS AND PRODUCTS INCORPORATING AI (SEPIA)

OBJECTIVE

Develop evaluation methods tailored to the cybersecurity of AI systems and cybersecurity products incorporating AI, intended for the ecosystem in order to strengthen the security of artificial intelligence systems and to provide a harmonised response to the various regulatory frameworks concerned (AI Act, Cyber resilience act, EU cybersecurity act). The aim of this work is also to promote and highlight these efforts within European and international bodies dealing with cybersecurity certification.

METHODOLOGY

The project is supported by an inter-agency working group (ANSSI, Inria, LNE, AMIAD, Information Technology Security Evaluation Facility - ITSEF) which meets monthly to develop the desired deliverables.
Depending on the opportunity, the work could include practical cases for evaluating the evaluation methods developed, through experiments involving ITSEF, INESIA parties and industrial players. A single product could be tested by several entities, or several products could be evaluated simultaneously.
Depending on the opportunity, awareness-raising activities could also be conducted with market surveillance authorities in application of the AI Act.
The project may be conducted in conjunction with Project No. 5, as necessary.

STAKEHOLDERS

ANSSI, Inria, LNE. ITSEF (Almond, CEA-Leti, Lexfo, Oppida, Quarkslab, Thales/CNES, Serma Safety & Security), AMIAD.

CONTEXT

In addition to the usual cybersecurity vulnerabilities, AI systems are exposed to specific vulnerabilities related to AI models. The impact of these vulnerabilities and the means of addressing them must be considered not only at the level of individual components, but also from the perspective of the AI system's architecture and its integration within an information system.
However, there is currently no evaluation and certification framework for evaluating the cybersecurity of AI systems. While the AI Act provides for compliance evaluations of AI systems, particularly with regard to cyber aspects, evaluation methods need to be developed to certify the level of cybersecurity of an AI system through evaluations and penetration tests carried out by a third-party assessor.
To address this, ANSSI has been leading the SEPIA (SEcurity of Products Integrating AI) project since early 2025, which aims to anticipate these changes by developing specific evaluation methods in collaboration with the French ecosystem and foreign partners.

DELIVRABLES

- Evaluation methods for the cybersecurity of systems and products incorporating AI;
- Inventory of known attacks (software, hardware), and identification of the effort required to implement them (level of vulnerability analysis).

SYSTEMIC RISKS PILLAR

Objectively assess and evaluate the systemic risks that AI could pose, inform public decision-making and assert France's position in AI security.

Technological advances and the deployment of AI are driving rapid changes in society and raising concerns about the systemic risks that could emerge. Embodying France's commitment to the AI Safety Institutes network, INESIA will work to inform public decision-making and contribute to a framework of trust conducive to the dissemination of AI.

By defining measurable thresholds and specifically modelling critical scenarios, Project No. 5 aims to **objectify systemic risks** related to large-scale information manipulation (health, financial system, information systems), offensive cyber security and CBRN proliferation. It will also propose mitigation methods.

Further downstream, Project No. 6 focuses on **evaluating agent systems** by developing protocols to test their reasoning, interaction and convergence capabilities in critical scenarios.

Project No. 7 affirms France's place in **the network of AI Safety Institutes**. By participating in its work, particularly in joint evaluations, France will confirm its place in global efforts to ensure AI safety and strengthen its own expertise. While the current wave of work is mainly embodied in Project No. 6, the scope will be redefined as needed in line with the programme for future waves.

PROJECT N°5.

DEVELOPING TECHNICAL EXPERTISE IN SYSTEMIC RISK EVALUATION AND MITIGATION: INFORMATION MANIPULATION, OFFENSIVE CYBERSECURITY, CBRN

OBJECTIVE

The project is part of a research initiative to assess the impact of AI in the field of systemic risks. The central issue here is to understand how AI models with high-impact capabilities can have real negative effects on public health, safety, public security, fundamental rights or society as a whole, which can be propagated on a large scale throughout the value chain.
The project will be based on an analysis of critical technological and behavioural thresholds that are likely to amplify or trigger systemic effects. It will aim to design evaluation and early detection protocols that can be used both by developers (to enhance model security) and by regulators (to anticipate and regulate risky uses).
Finally, the project aims to provide public decision-makers with the ability to anticipate and gain insight into the future, based on vulnerability indicators and impact scenarios. This support will contribute to the development of prevention, supervision and governance policies adapted to the rapid evolution of AI capabilities and uses.

CONTEXT

Certain uses of AI could have far-reaching effects that are difficult to predict and could be exploited for malicious purposes: disinformation and cyber security attacks, proliferation, destabilisation of an economic system, etc. Technological advances are lowering the barriers to entry for the development of models with a high potential for harm, whether deliberate or accidental.
In the face of these threats, current evaluation frameworks remain insufficient. It is becoming necessary to better understand how to assess the risks associated with the use of AI, to identify the thresholds of capability beyond which a system can become dangerous, and to formalise recommendations for auditing and containing them.

METHODOLOGY

The project mainly involves fundamental and applied research activities. Organised around three complementary themes, these activities aim to develop AI evaluation methodologies to characterise, test and anticipate distinct forms of systemic risk associated with advanced AI models, while developing mitigation approaches:

- Assessing the capacity of AI models to contribute (intentionally or unintentionally) to risks, particularly in the CBRN field. In this latter field, conducting uplift studies, identifying critical thresholds, and designing test benches to detect when these thresholds are crossed;
- Analyse the persuasive and manipulative capabilities of conversational agents on human users, to better understand the conditions under which mass manipulation emerges and to design vigilance and alert tools. For example, studying the persuasive power of conversational agents on humans, detecting attempts at mass manipulation by identifying the content generated, the methods used to generate it, and the campaign itself;
- Analysing the capabilities of cyberattacks, given that AI can be used to generate actions targeting the confidentiality, integrity or availability of information systems. As the exact level of threat has not yet been rigorously assessed, further research is needed to answer various questions of interest.

In parallel with this evaluation work, the project will develop mitigation methods. In particular, it will propose safeguards integrated into the AI systems themselves (filtering of training or fine-tuning data, training and optimisation recommendations, abuse detection, internal consistency checks), and will seek defensive countermeasures to strengthen the resilience of infrastructures and applications targeted by malicious uses of AI.

The project may be carried out in conjunction with other projects as necessary, in particular Project No. 4.

DELIVRABLES

- Identification of research projects specifically focused on systemic risk evaluation and mitigation;
- Publication of mitigation methods associated with open source tools, where appropriate;
- Creation of non-public evaluation sets, where appropriate.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN. VIGINUM and AIST (SGDSN) as external partners.

PROJECT N°6.

ASSESSING THE PERFORMANCE AND RISKS OF AGENTIC SYSTEMS

OBJECTIVE

Develop a rigorous methodology for evaluating AI agent systems based on large language models. In particular, characterise their capabilities in terms of cybersecurity and assess the extent to which they could be used for criminal purposes, with a view to informing public authorities and refining risk anticipation. Contribute to the work carried out within the AI Safety Institutes network in order to assert France's position within the network.

CONTEXT

Agent architectures are now one of the main drivers of AI capacity building and a major vector for its widespread adoption. In 2024, the ecosystem saw the emergence of a new generation of agentic systems, combining LLMs, external tools, long-term memory and reasoning or planning capabilities. Capable of breaking down a task, mobilising tools and interacting with a complex environment, they pose unprecedented challenges:

- the final result only partially reflects the quality of the underlying reasoning;
- behaviours are highly dependent on experimental conditions;
- certain safety-critical uses must be anticipated through targeted scenarios.

METHODOLOGY

The project will begin with a state-of-the-art literature review of identified risks and security-oriented evaluation criteria and protocols. This state-of-the-art review may be enriched by use cases provided by the project's stakeholders. The list of identified risks may also be expanded to include additional use cases identified by the project partners. The objective is to list, on the one hand, the types of potential risks through critical scenarios and, on the other hand, the evaluation criteria and protocols selected for INESIA.

In a second phase or through successive iterations, these protocols will be implemented, drawing on existing resources where relevant. This will involve reviewing existing evaluation libraries, developing new ones where necessary, and using or creating dedicated reference datasets.

DELIVRABLES

- Report on state-of-the-art security-oriented risk and evaluation protocols, enabling the selection of certain critical scenarios;
- Identification or development of an evaluation library that implements these protocols;
- Specifications, then identification or creation of evaluation datasets.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN.

PROJECT N°7.

PARTICIPATING IN THE ACTIVITY OF THE NETWORK OF AI SAFETY INSTITUTES

OBJECTIVE

Contribute to the work carried out within the AI Safety Institutes network, in order to strengthen INESIA's expertise in the field of evaluation and assert France's position within the network.

CONTEXT

The AI Safety Institutes network facilitates international cooperation in the field of AI evaluation. At the AI Action Summit organised by France on 10 and 11 February 2025, representatives from several parts of the international AI Safety Institutes (AISI) network gathered in Paris to unveil the initial results of joint testings. France provided the dataset used to assess the cyber robustness of the models and carried out performance evaluations relating to security risks when interacting with the tested models in French.
In addition, the work carried out during the first three waves of evaluation were showcased at the ICML conference in Vancouver. In line with the corresponding project in this roadmap, this work focused on the evaluation of agentic systems.

METHODOLOGY

By nature, the precise content of this project will remain to be defined in accordance with the guidelines set within the AI Safety Institutes network. INESIA members may propose certain topics related to their activities, which the SGDSN may bring to the attention of the network when defining the programme for future waves of work.

DELIVRABLES

- Joint evaluation work or other activities, depending on the network's work agenda;
- Where relevant, scientific publications and communications.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN.

PERFORMANCE AND RELIABILITY PILLAR

Stimulate creativity within the ecosystem, using emulation as a lever to encourage the emergence of more efficient and reliable solutions.

This third pillar of INESIA will contribute to improving the reliability and performance of AI models and systems, insofar as considerations of general interest justify the involvement of public authorities.

Organising challenges is a key lever for stimulating the ecosystem around topics of interest for AI evaluation or security. By fostering a spirit of 'coopetition', INESIA would be able to bring about creative solutions. Project No. 8 invites proposals for ideas and frameworks for such challenges.

PROJECT N°8.

ORGANISATION OF CHALLENGES

OBJECTIVE

The challenges aim to stimulate international competition, enabling rapid maturity of AI technologies through an internationally open evaluation campaign. The challenge is to create a significant ripple effect to clarify and advance the state of the art. The aim is to foster 'coopetition' between the participating consortia, creating a unifying and structuring emulation that promotes exchanges between experts, ensuring their mobilisation and maximising innovation and technological progress.

CONTEXT

The evaluation of AI faces many technical and methodological challenges in order to enable its widespread and confident adoption in businesses and society: reliability, multimodal management, content personalisation, adaptation to a specific context or terminology, etc. Organising challenges allows the performance and reliability of AI to be verified in a collaborative and comparative manner, enabling both the science of evaluation to advance based on concrete cases provided by challenge participants and innovation to be stimulated.

METHODOLOGY

The challenges will be structured in such a way as to encourage international competition and enable different players in the AI ecosystem to come together to solve a specific artificial intelligence evaluation task. In order to attract a diverse range of participants, both French and international, the challenges must focus on operational issues, be easy to implement and enable concrete results to be achieved quickly. A section devoted to scientific outreach (Project No. 11) could also serve as a vehicle for communicating and disseminating INESIA's work. The details of the selected challenge will be specified in consultation with the INESIA strategic committee when the specifications are drawn up.

DELIVRABLES

The initial work on this project will focus on organising a pilot challenge, which will serve as the basis for subsequent INESIA challenges.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN.

CROSS-CUTTING AXIS

Create structuring commons for INESIA's work, develop synergies between the parties and their work, and bring together an ecosystem.

Beyond the activities carried out within the three clusters, a range of work is necessary to ensure the smooth running and influence of INESIA. By establishing a clear framework and common tools, and by sharing knowledge and best practices, the parties will strengthen their synergies and accelerate their growth in expertise.

The primary challenge is to capitalise on current knowledge and supplement it with a comprehensive body of knowledge on AI evaluation. Project No. 9 will aim to identify available methods, tools and concepts, pinpoint gaps, and provide political authorities with a clear and dynamic summary of the knowledge that can be mobilised.

It is also necessary for INESIA to have evaluation resources that meet its needs, in particular those capable of supporting the evaluation of sensitive models and participating in coordinated exercises led by AI Safety Institutes (Project No. 10).

Finally, a series of communication and horizon scanning activities will highlight INESIA's work and deepen discussions on AI evaluation and safety at the national and even international level. By facilitating exchanges on a broad scale encompassing both private and public ecosystems, disseminating best practices and highlighting INESIA's work, it will help to structure the community (Project No. 11).

PROJECT N°9.

ESTABLISHING A FRAMEWORK FOR COORDINATING ACADEMIC AND METHODOLOGICAL MONITORING ACTIVITIES

OBJECTIVE

Several of INESIA's projects are based on monitoring activities: Project No. 1 (Providing evaluation tools to national regulators) and Project No. 10 (Ensuring access to an AI evaluation solution) require monitoring of methodological advances in evaluation, while Project No. 11 (Scientific coordination of INESIA's work) involves monitoring academic publications. Project No. 9 will aim to centralise, standardise and coordinate the knowledge acquired by INESIA partners during the execution of their work.
These activities will provide a central resource for AI evaluation and offer support for strategic decisions concerning research and development efforts dedicated to AI evaluation.

METHODOLOGY

The project will propose a framework and technical basis for centralising and linking the results of review and monitoring work, both academically and methodologically. These state-of-the-art reviews and monitoring may concern, for example, knowledge acquired relating to benchmarks, metrics, evaluation methods and tools. This work will be based on analysis of the literature and results produced by the scientific and industrial community.

CONTEXT

While there is currently no consensus on methods for evaluating advanced AI models and systems, the work of the AISI network has highlighted significant discrepancies between evaluations carried out by different countries, even for models evaluated under similar conditions.
The activities carried out by INESIA require state-of-the-art work in a variety of fields. To date, the parties have already established knowledge bases that cannot be exploited by others because they have not yet been shared. In order to maximise their usefulness and provide INESIA with a genuine reference framework, it is necessary to establish a framework that brings this knowledge into line. Where necessary, links with other projects, in particular Project No. 1, Project No. 10 and Project No. 11, will be considered.

DELIVRABLE

- Pooling of monitoring work within INESIA.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN.

PROJECT N°10.

ENSURING ACCESS TO AN AI EVALUATION FRAMEWORK & INFRASTRUCTURES

OBJECTIVE

Design, develop and deploy software and hardware infrastructure suitable for evaluating AI systems based on use cases provided by INESIA. This modular platform will be capable of conducting reproducible evaluations within the framework of interoperable protocols, ensuring secure testing on sensitive models and enabling contributions to international experiments on AI evaluation (e.g. AI Safety Institutes, European Commission AI Office).

METHODOLOGY

The first step is to specify the use cases in the context of INESIA and the current practices of the actors involved in the various projects. Next, an evaluation of open-source frameworks for evaluating generative models will be carried out, along with the tools needed to interface with experimenters and to orchestrate and monitor experiments.
Based on a comparison of the identified needs and existing capabilities, this study could lead to the production of initial specifications for the platform required to meet INESIA's needs. In this regard, the possibility of contributing to the open-source ecosystem will be explored if a link with INESIA's activities is proven relevant.
The work in this project would be structured around three elements:

- the interface with users (experimenters);
- the orchestration and monitoring of experiments, with the addition of customised software components;
- the computing resources made available.

CONTEXT

The evaluation of AI systems requires software and hardware infrastructure tailored to the needs of INESIA, as defined in this roadmap. These needs could include, but are not limited to, the following use cases: evaluation of general-purpose AI models and autonomous agents, participation in AISI work where appropriate, design and regular updating of the basic components (datasets, evaluation tools, replication scripts) necessary to maintain reliable leaderboards. In particular, the platform will need to meet the specific requirements of model evaluations that require a high level of confidentiality, whether they are non-public or high-risk. It will be able to host private evaluation games necessary for the proper conduct of activities.
As mentioned, and as necessary, links with other projects, in particular Project No. 6, Project No. 7, Project No. 9 and Project No. 11, will be considered.

DELIVRABLES

- Determination of INESIA's needs, identification of available capacities and, based on the proposed analysis and available resources, design of a software and hardware infrastructure for launching experiments with AI models and agents.

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN. External partners may be involved.

PROJECT N°11.

SCIENTIFIC EVENT HIGHLIGHTING ACADEMIC AND INDUSTRIAL CONTRIBUTIONS TO ADVANCING THE SCIENCE OF EVALUATION

OBJECTIVE

Structure, lead and unite the national and even international community around the theme of AI evaluation. Accelerate the maturation of evaluation methodologies, promote the sharing of best practices, stimulate the comparison of approaches, and enhance the visibility and attractiveness of French research in this strategic field. Facilitate monitoring of scientific topics of interest to INESIA.

CONTEXT

The evaluation of AI faces major methodological obstacles: lack of consensus on metrics, diversity of use cases, uncertainty about the properties to be measured. INESIA must therefore bring together academic, industrial and institutional expertise in order to discuss the state of the art.

METHODOLOGY

INESIA will organise annual scientific conferences to share the various methodological and technological advances in AI evaluation. To this end, it will monitor methodological developments in order to identify the latest advances in this field. These advances will be listed in a bibliographic monitoring document and the most important ones will be shared with INESIA's various partners. The results of the challenges may also be presented there.
Calls for papers may be opened and will cover a wide range of topics: methodological foundations, feedback, open-source tools, sector benchmarks, etc.
By its very nature, this project is likely to interface with other projects in order to highlight their outputs.

DELIVRABLES

- Conference proceedings (papers, posters, recommendations, etc.)

STAKEHOLDERS

ANSSI, Inria, LNE and PEReN. The academic community, start-ups, industries and challenge platforms may be involved.