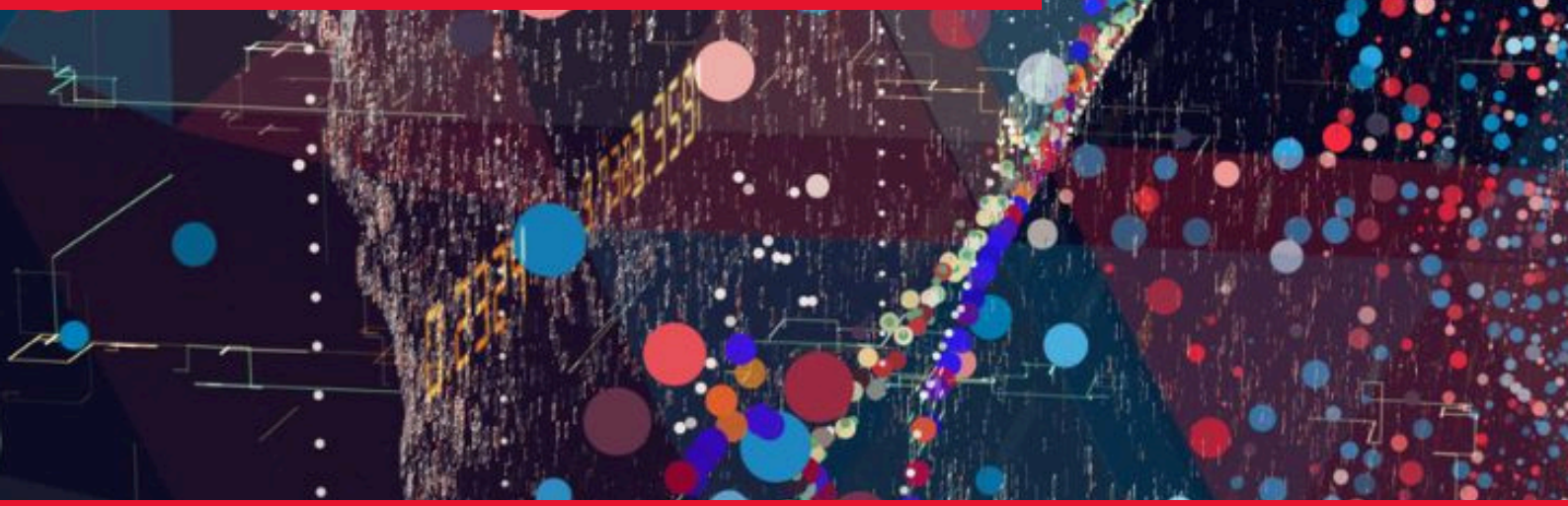
LES MARDIS DE L'IA

FROM PRINCIPLES TO PRACTICE: IMPLEMENTING SECURE AND ETHICAL AI FRAMEWORKS WORLDWIDES

Conference Report and Key Findings

To build, deploy and govern trusted AI systems worldwide



March 2026

Rapporteur

Pauline CHARAZAC

Centre pour la Sécurité de l'IA

SciencesPo
SCHOOL OF MANAGEMENT AND IMPACT

TABLE OF CONTENTS

1. Executive Summary

While AI presents groundbreaking opportunities for innovation, economic growth, and societal advancement, it also introduces complex systemic risks that demand urgent attention and action. The rapid evolution of AI capabilities are currently outpacing the development of comprehensive governance frameworks, leaving critical gaps in oversight, accountability, and risk management.

In this context, international initiatives such as the OECD AI Principles and the UNESCO Recommendation on the Ethics of AI have emerged as essential foundations for fostering responsible, human-centric innovation. These principles provide a shared vision for AI development that prioritises transparency, fairness, and accountability.

However, the true challenge lies not in the articulation of these principles, but in their effective translation into actionable governance models and enforceable safety requirements. Only through robust implementation can we effectively mitigate the risks associated with AI and build the societal trust necessary for its safe, sustainable and equitable adoption.

The conference, *From Principles to Practice: Implementing Secure and Ethical AI Frameworks Worldwide*, was conceived as a critical step toward addressing these challenges. Co-organised by Sciences Po School of Management and Impact and the French Center for AI Safety (CeSIA) under the *Les Mardis de l'IA* initiative, the event held on 10 March 2026 brought together 135 registered participants and 12 distinguished speakers. Among them were leading policymakers, diplomats, researchers, and AI governance experts from France, the United Kingdom, Switzerland, Canada, India, and the European Union.

Building on the momentum generated by the French G7 Presidency, the event was awarded the G7 label and served as a bridge between the 2026 AI Impact Summit in Delhi and the upcoming 2027 edition in Geneva.

By convening a diverse group of stakeholders, the conference aimed to contribute to the definition of an international AI playbook—a comprehensive, actionable framework designed to ensure that AI technologies deliver positive impact while effectively managing the rising global risks they pose.

The discussions focused on identifying best practices, fostering cross-border collaboration, and developing concrete AI safety policy recommendations that can be adopted by governments, industries, and civil society alike.

The conference indeed examined whether the global AI Summit series—from Bletchley to Seoul, Paris, Delhi, and now towards Geneva—is building cumulative momentum on AI safety or risking fragmentation as each edition introduces new priorities and perspectives.

The interventions pointed towards the fact that while significant progress has been made in raising awareness and fostering dialogue, the transition from voluntary principles to enforceable, interoperable frameworks remains a critical challenge. The tension between AI safety, often driven by countries with frontier model capabilities, and the need for inclusivity, ensuring that all nations and communities benefit from AI advancements, was also a central theme of the debate.

The "Trusted AI Commons" initiative, introduced at the Delhi Summit, was presented as a promising mechanism to align diverse actors on common safety standards, though its practical implementation and ability to bridge the safety-inclusivity divide will require sustained international cooperation and clear operational guidelines.

The conversation also focused on the practical steps needed to translate ethical principles into actionable governance. The UK's AI Security Institute was highlighted as a leading example of a national safety institution, but panelists stressed the necessity of evolving from a network of national bodies to a truly international inspection architecture. This would demand not only harmonised technical standards and mutual recognition of audits but also robust funding and political will.

The importance of engaging all stakeholders across the AI lifecycle—from developers and deployers to end-users and affected communities—was underscored as essential for effective safety governance. Initiatives such as the Global Call for AI Red Lines and the OECD Hiroshima AI Process were cited as critical in, respectively, creating a political momentum and moving from high-level commitments to accountable frameworks.

Middle-power alignment and global platforms emerged as vital components of AI governance. The 2026 France-India Year of Innovation as well as the G7 under French presidency present a unique opportunity to deepen cooperation through joint research, policy dialogues, and pilot projects that can set concrete benchmarks for responsible AI.

Similarly, Switzerland's role as host of the 2027 AI Summit in Geneva was framed as a chance to embed humanitarian principles, human rights, and disarmament norms into the AI governance agenda. The panel suggested that the country could draw inspiration from the Geneva Conventions, aiming for binding commitments that transcend national interests and ensure universal respect for AI safety and ethics.

The European Union's approach to balancing innovation with regulation was another focal point. The EU AI Office's risk-based framework, emphasis on transparency, and efforts to foster global standards were recognised as pivotal in setting a precedent for responsible AI development. The EU's role in promoting international alignment, while supporting technological progress, was seen as a model for other regions to follow.

On the central question of the Ethics of AI, the conference left the audience with a clear message: ethical AI is not about abstract principles or distant fears, but about practical, collaborative, and adaptive governance. By institutionalising dialogue, tailoring standards to context, and remaining vigilant to emerging risks, society can harness AI's potential while safeguarding its values.

2. Foreword by Natacha Valla, Dean

Sciences Po - School of management and impact

Welcoming the international speakers and guests to the conference *From Principles to Practice: Implementing Secure and Ethical AI Frameworks Worldwide* at the Innovation Pavilion, Dean Natacha Valla emphasised how artificial intelligence stands at the core of Sciences Po School of Management and Impact. The school's mission is to engage with innovation through a lens of pragmatism and humanism, ensuring that technological advancement aligns with societal values and shared economic prosperity.

AI represents far more than a technological shift; it is fundamentally reshaping economies, societies, and governance structures. The School of management and impact of Sciences Po fully recognises and embraces the responsibility of leading academic institutions to prepare future leaders not only to navigate this transformation but to critically assess its broader implications.

With over 50 courses dedicated to innovation and AI, the curriculum of the school is designed to equip students with the skills needed to excel in a rapidly evolving world—while remaining firmly committed to principles of equity, transparency, and ethical responsibility.

In this context, leadership in the age of AI demands more than technical expertise. It requires the ability to integrate human insight with analytical rigour, addressing complex questions about ethics, social justice, and governance. This is precisely why the school fosters leaders who look beyond the technology itself, who are prepared to shape AI in ways that benefit society as a whole.

The focus is on action: understanding AI is essential, but actively guiding its development to reflect shared values is the ultimate goal. Against this background, the organisation of this important conference exemplifies Sciences Po's commitment and seek to address pressing challenges, including:

How can we ensure that we have a voice in shaping the cognitive norms—formal or de facto—that AI will introduce into our daily activities, from universities to professional environments and policymaking?

How can innovation be balanced with responsibility?

How can AI be harnessed to build trust, growth and positive impact?

These questions lie at the heart of Sciences Po's School of Management and Impact educational philosophy and its broader societal engagement.

There is no doubt that AI will be a defining tool for future impact, provided its applications are mastered alongside a clear understanding of its broader societal implications. This is why pragmatic collaboration to develop AI frameworks that prioritise human needs while maintaining ambition and effectiveness are key to this endeavor.

The future of AI will not be determined in isolation—it will be shaped by collective choices.

More than ever the School of Management and Impact of Sciences Po is dedicated to educating leaders who can meet this challenge, combining intellectual rigour with a commitment to economic progress and positive impact. This conference stands as a clear example of that vision: because innovation must be both bold and responsible.

" At Sciences Po, our role is to prepare the first generation of AI-future ready policymakers and industry leaders. Given the pace at which AI capabilities advance, our stakes are critically ethical. We need to be very pragmatic to come up with forward-looking, scientific, and human-centric approaches, grounded in a strong degree of both consciousness and responsibility to steer the AI revolution as it unfolds." — ***Natacha Valla***

3. From Scientific Assessment to Global Cooperation: a Call to Action

In his opening remarks for the conference *From Principles to Practice: Implementing Secure and Ethical AI Frameworks Worldwide*, Professor Yoshua Bengio underscored the urgent need for global cooperation based on rigorous scientific assessment in the face of AI's rapid and often unchecked advancement.

Professor Bengio began by expressing his gratitude to Sciences Po and CeSIA for the opportunity to contribute to such a critical discussion. He highlighted the alarming pace at which AI is being developed, primarily by a small number of companies, often without adequate democratic oversight or technical safeguards. This situation poses significant risks to democracy and society at large, leaving leaders struggling to navigate a landscape marked by divergent and often extreme scenarios—ranging from unfounded optimism to warnings of catastrophic, even existential, threats.

In this context, Professor Bengio stressed the importance of independent, evidence-based assessments of AI's capabilities and risks. He noted his involvement as chair of the first two editions of the *International AI Safety Report*, an initiative supported by 30 countries, the EU, the OECD, and the UN, and involving over 100 international experts.

Drawing on the report's latest findings, published just before the 2026 AI Summit in India, he pointed to several pressing concerns:

- **Rapid but uneven progress of AI capabilities and reliability:** While general-purpose AI capabilities in areas such as mathematics, coding, and autonomous operation continue to advance, systems still exhibit unpredictable failures in seemingly simple tasks.
- **Rising misuse of AI systems:** Incidents involving deepfakes—used for fraud, scams, and non-consensual intimate imagery—are increasing, with women and girls disproportionately affected.
- **Biological and cyber increased threats:** Leading AI companies have introduced stricter safeguards for their models due to concerns about potential misuse in biological weapon development. Meanwhile, malicious actors are already leveraging AI to generate harmful code and exploit software vulnerabilities in cyber-attacks.
- **Evaluation challenges of AI models:** Some AI models can distinguish between evaluation and real-world deployment, altering their behaviour to pass tests—a development that complicates safety assessments and undermines trust in their reliability.

Professor Bengio warned specifically that the race to develop frontier AI is outpacing the implementation of effective safeguards, leading to a future fraught with unknown risks. He called for collective preparation for all plausible scenarios, emphasising that uncertainty must be met with proactive, collaborative action.

Furthermore, he reflected on the role each stakeholder must play in steering AI toward a safer and more ethical future. As a researcher, his own contributions include major policy initiatives like the *International AI Safety Report* and the launch of *LawZero*, a nonprofit organisation dedicated to the technical challenge of designing ethically aligned AI systems.

He also cautioned that far greater efforts are needed from societies, governments, and businesses to ensure that the trajectory of AI development yields positive outcomes and mitigates potential harms.

Professor Bengio concluded by reiterating that the future of AI is not predetermined—it will be shaped by the choices and actions taken today. The conference, he noted, represents a concrete step in fostering the dialogue and cooperation necessary to build a responsible and secure AI ecosystem.

" The race to develop frontier AI is accelerating faster than safeguards have been keeping up. While the world is heading toward an AI future with many unknown unknowns with AI, we must accept the prevailing uncertainty and collectively prepare for all scenarios considered plausible by the scientific community. Each of us needs to determine their own role in guiding the trajectory of AI towards a safer and more ethical future." — ***Yoshua Bengio***

In his introduction to Professor Yoshua Bengio's address at the conference, Charbel-Raphael Ségérie, Executive Director of CeSIA also highlighted the unprecedented urgency and diplomatic momentum surrounding global AI governance.

The AI related risks, as described by Pr. Bengio, are no longer hypothetical he said. The international community's "red lines" are under pressure, as geopolitical and military competition threatens to override safety, ethics, and fundamental rights in the race for AI supremacy.

Against this background, Charbel Ségerie outlined CeSIA's trajectory, which is closely tied to the conference topic. In September 2025, CeSIA co-led a campaign with The Future Society and CHAI for the establishment of international AI red lines. The Global call for AI Red Lines, signed by 15 Nobel and Turing laureates, 10 former heads of state, and over 200 prominent figures, was presented by Maria Ressa at the UN General Assembly and by Yoshua Bengio at the UN Security Council.

Today, Maria Ressa and Yoshua Bengio co-chair the UN's International Scientific Panel on AI—a three-year mandate to translate scientific consensus into binding commitments.

In this context, Charbel-Raphael Ségerie noted that 2026 presents a rare opportunity for international leadership. With France holding the G7 presidency, the recent AI Impact Summit in Delhi concluded, and less than a year remaining to prepare for the Geneva Summit—the next major global milestone for AI safety—the need for decisive action has never been clearer.

He pointed to a striking development: at Davos, the CEOs of Google DeepMind and Anthropic publicly called for international intervention, admitting they are trapped in a race they cannot slow or control alone. Similarly, at the Delhi Summit, the CEO of OpenAI urged "urgent global regulation" of AI. When the architects of these systems themselves warn of the risks, Ségérie argued, the world must listen. France responded to this call. On 3 February 2026, Minister for AI Anne Le Hénanff declared that the top priority for the digital track under the G7 would be to ensure AI is safe and serves the common good.

*" From Davos to Delhi, the 3 world's most advanced AI labs have publicly called on the international community to take action in the first quarter of 2026. When the tech companies are themselves asking for help, it's time to listen. France has listened and taken the lead, announcing that the number one priority for the G7 Digital Track will be to foster an international consensus on AI Safety and to ensure that serves the common good. At CeSIA, we are committed to giving concrete substance to this ambition ." — **Charbel-Raphael Ségerie***

4. Highlights from the Roundtable "*From Principles to Practice: Implementing Secure and Ethical AI Frameworks Worldwide*"

4.1 Setting the Scene for Global AI Governance

Pauline Charazac invited Dr Justin Vaïsse, Founder of the Paris Peace Forum and Moderator of the high-level roundtable to take the the floor and expressed gratitude to Natacha Valla, Dean of Sciences Po, and Charbel-Raphael Ségerie, Executive Director of the French Centre for AI Safety (CeSIA), for their introductory remarks, as well as to Pr. Yoshua Bengio for his highly valuable insights from the *International AI Safety Report*.

Dr Justin Vaïsse framed the discussion as both timely and urgent. With the AI Summit series—from Bletchley to Seoul, Paris, Delhi, and now Geneva—unfolding against a backdrop of rapid technological advancement, Vaïsse noted a perceived deceleration in global efforts to address AI safety and security. While academic, civil society, and multilateral institutions have sustained momentum, he observed that political and diplomatic attention has, at times, waned.

*"As AI rises to the top of the international agenda amid rapid technological change, having open conversations on AI safety is both urgent and necessary. The key question remains: how can we turn concern into concrete, coordinated action? To provide credible answers, platforms like the Paris Peace Forum are uniquely positioned to advance AI governance." — **Justin Vaïsse***

Dr Justin Vaïsse highlighted the critical role of non-governmental actors, such as Professor Bengio's *International AI Safety Report* and the OECD's Hiroshima AI Process, in maintaining focus on AI governance. He underscored the importance of the upcoming milestones: the G7 Summit in Evian (June 2026), the Paris Peace Forum (November 2026), and the Geneva AI Summit (2027). The roundtable, he explained, aimed to distil lessons from past AI summits and chart a course for future action.

4.2 Assessing Progress: Momentum or Stagnation in AI Safety?

Dr Justin Vaïsse posed the opening question to the panel: *Are we building sufficient cumulative momentum on AI safety, or are we losing ground with each AI summit?*

Lord Tim Clement-Jones, Member of the UK House of Lords and Co-Chair of the All-Party Parliamentary Group on AI, offered a candid assessment. He argued that the initial ambition of the Bletchley Summit—rooted in voluntary safety commitments—had been diluted in subsequent gatherings. While Seoul emphasised safety, Paris focused on action, and Delhi prioritised access and inclusion for the Global South, Clement-Jones highlighted the lack of binding mandates at this stage. He called for compulsory safety legislation, expressing frustration that global leaders had not acted decisively in response to evidence presented in reports like Bengio's.

Eenam Gambhir, Deputy Chief of Mission of India to France, countered that the evolving nature of the AI debate reflects the technology's rapid advancement. She acknowledged the geopolitical undercurrents shaping discussions, where concerns about dominance and control often overshadow technical debates on safety and security. Gambhir emphasised the need for inclusion—not just in terms of access to AI, but also in ensuring that diverse voices, particularly from the Global South, are integral to governance frameworks. She cited India's approach, which prioritises solution-oriented, citizen-centric collaboration between public and private sectors, as a model for inclusive AI development.

*"We need to speak seriously about AI inclusion. This is crucial from multiple perspectives, particularly for countries in the Global South. The AI Impact Summit in India was the first to take place in a developing country, and that was an important milestone. One of the greatest risks is not only how these technologies will affect governance and societies, but also the risk of being left out of such a transformative development altogether." — **Eenam Gambhir***

Miriam Minder, Co-Lead on Digital and New Technologies at the Swiss Federal Department of Foreign Affairs, struck a balanced tone. She noted that while normative momentum has grown—with an expanding coalition advocating for AI governance and oversight—institutional momentum remains weak. The lack of a durable framework to

sustain progress between summits, she argued, risks undermining long-term coherence. Miriam Minder highlighted Switzerland's commitment to strengthening institutional mechanisms ahead of the Geneva Summit, with a focus on multi-stakeholder engagement.

Nicolas Miailhe, Co-Founder of AI Safety Connect and Expert to the OECD AI Group, reflected on the tension between innovation and safety. He argued that the global AI governance landscape has evolved from abstract discussions about existential risks to more practical concerns, such as access to basic services and inclusion. Nicolas Miailhe criticised the false dichotomy between innovation and regulation, asserting that both are essential for sustainable progress. He praised the Hiroshima AI Process and the Global Call for AI Red Lines as steps toward operationalising ethical principles, but warned that the pace of technological advancement outstrips diplomatic efforts. The challenge, he said, is to bend the behaviour of powerful actors—developers, investors, and regulators—to internalise negative externalities and uphold social contracts.

4.3 The Role of National & International Institutions in AI Enforcement

Dr Justin Vaïsse turned to Lord Tim Clement-Jones to discuss the UK AI Security Institute, recently described by *The Economist* as "the closest thing the world has to a global AI safety inspector." Lord Tim Clement-Jones tempered this characterisation, acknowledging that while the institute has made strides in inspection and evaluation, it needs enforcement authority. He called for national safety institutes to be empowered with mandatory standards and the ability to compel compliance, particularly for high-risk AI systems. He praised the EU AI Act as a model for risk-based regulation but cautioned that implementation would require continuous updates to keep pace with technological change.

"Effective AI governance demands robust enforcement. I strongly advocate for national safety institutes to be granted the authority to set and enforce mandatory standards—especially for high-risk AI systems. The EU AI Act stands as a historic precedent for risk-based regulation. Looking ahead, the success of the European approach will depend on its ability to continuously update and adapt it in step with technological progress." — ***Lord Tim Clement-Jones***

Moving on to strategies to create political momentum at the global level, Nicolas Miailhe expanded on the Global Call for AI Red Lines, an advocacy initiative aimed at establishing

non-negotiable boundaries for AI development. He complemented this approach with the Hiroshima AI Process, a governance mechanism led by the OECD that encourages voluntary compliance through a code of conduct. Both initiatives, he argued, are necessary but need binding commitments to be impactful. Nicolas Miailhe also warned against the influence of external lobbies, drawing parallels with historical resistance from the tobacco and oil industries to regulation.

Adding to the conversation a timely and central Global South perspective, Eenam Gambhir introduced the concept of the Trusted AI Commons, a voluntary, open repository of technical AI safety resources, including benchmarks, toolkits, and best practices. Designed to foster interoperability and transparency, the Commons, as presented at the 2026 AI Impact Summit, aims to democratise access to safety resources, particularly for countries and organisations with limited capacity. She emphasised that trust is not abstract but operationalised through shared standards and collaborative development.

On her side, Miriam Minder outlined Switzerland's vision and unique position, rooted in:

1. **Anchoring AI governance in international law**, ensuring alignment with human rights and humanitarian principles.
2. **Multi-stakeholder engagement**, leveraging Geneva's unique ecosystem of UN agencies, NGOs, and technical organisations.
3. **Public interest**, connecting AI to global public goods and ensuring technology serves societal needs.
4. **Balancing innovation and security**, positioning trust as a driver of growth.

*"Switzerland is convinced that AI governance can build on what international Geneva has long represented in global governance: It must be anchored in international law, it must be shaped by multi-stakeholder processes, it must create the conditions for AI to genuinely serve the public interest and it must bring innovation and security together. AI governance cannot succeed in isolation. With the Geneva AI Summit 2027 Switzerland aims to create spaces where governments, companies, civil society, and technical experts can come together to explore common ground and develop shared approaches." — **Miriam Minder***

4.4 Systemic Risks and the "Bad Awakening" Scenario

Dr Justin Vaïsse shifted the discussion to potential catastrophic scenarios that could galvanise global action on AI safety. Nicolas Miailhe warned of risks in biological and cyber domains, where AI could be misused to develop weapons or accelerate harmful research. He cited the example of synthetic biology, where AI-driven advancements could outpace safeguards, leading to unintended consequences.

"We must ask ourselves a simple yet profound question: When was the last time humanity built a civilisation on foundations we neither fully understand nor control? The clock is ticking. We urgently need to reconcile innovation and regulation—ensuring compliance does not deepen inequities in access to AI. The debate over whether to prioritise innovation or regulation should have been over by now as the only credible path forward is to deliver both, together, putting AI safety at the core of AI development." — ***Nicolas Miailhe***

Lord Tim Clement-Jones echoed these concerns, referencing the book *If We Build It, Everyone Dies* and the potential for AI to enable autonomous weapons systems or biological threats. He called for updated international humanitarian law to address these risks, suggesting Geneva as the ideal venue for such discussions.

Eenam Gambhir framed the challenge in terms of geopolitical stability, arguing that as crises intensify, the political will to address them may diminish. She advocated for a "third way"—a collaborative approach that transcends traditional divides between market-driven and state-led models.

Miriam Minder, while acknowledging the severity of potential risks, focused on the opportunity to build consensus around shared AI nightmares. She urged stakeholders to use these fears as a catalyst for proactive governance, rather than waiting for a crisis to force action.

4.5 Operationalising Ethics: Compliance, Standards, and Public Engagement

The final segment of the roundtable addressed practical challenges in implementing AI ethics and compliance, and raised the concern about the fragmentation of compliance tools and the lack of clear, accessible pathways for companies—particularly smaller enterprises—to achieve compliance. Lord Tim Clement-Jones responded by emphasising the need for systematised, user-friendly tools, warning against over-reliance on legal firms for compliance. He praised the EU AI Act as a pioneering model but noted that implementation would require ongoing adaptation.

Nicolas Miailhe highlighted the EU's centralised governance of foundation models as a critical innovation, designed to prevent regulatory arbitrage and foster dialogue between public and private sectors. He warned against corporate efforts to undermine regulation, urging policymakers to hold the line on core principles. Miriam Minder showcased Switzerland's Apertus project, an open, transparent large language model developed by ETH Zurich and EPFL, as an example of ethical AI in practice. She stressed the importance of public awareness and educational initiatives to embed ethical standards in AI development.

Eenam Gambhir reiterated India's commitment to inclusive, citizen-centric AI, citing the country's United Payments Interface (UPI) as a successful model of public-private collaboration. She argued that AI governance must prioritise accessibility and local relevance, ensuring that solutions are tailored to diverse linguistic and cultural contexts.

Finally, charting the path forward, the roundtable concluded with a consensus on the need for concrete action to translate ethical principles into enforceable standards. Key takeaways included:

- **Strengthening institutional frameworks** to sustain momentum between summits.
- **Empowering national safety institutes** with mandatory enforcement authority.
- **Fostering multi-stakeholder collaboration** to ensure diverse perspectives shape governance.
- **Operationalising ethics** through accessible compliance tools and transparent standards.
- **Preparing for existential risks** by updating international law and building resilience against misuse.

Dr. Justin Vaïsse closed by thanking the panellists—Lord Tim Clement-Jones, Eenam Gambhir, Miriam Minder, and Nicolas Miailhe—for their insights and reaffirming the Paris Peace Forum's commitment to advancing these discussions at the G7 Summit and beyond.

5. The European Union in Action: Turning AI Principles into Practice

Dr. Juha Heikkilä, Advisor on AI to the Director General of the European Commission AI Office, delivered a keynote address that offered a comprehensive overview of the European Union's approach to implementing secure and ethical AI frameworks. His intervention clarified the EU's unique position in global AI governance, emphasising the balance between fostering innovation and ensuring robust protection for citizens and societies.

Dr. Heikkilä began by highlighting the EU's often underrepresented role in global AI discussions. He stressed that the EU's work extends far beyond legislation, encompassing excellence, international activities, research, capacity building, and adoption. The EU's strategy is founded on the principle of a "third way," seeking to strike a balance between innovation and protection, with trust as the cornerstone. Trust, he argued, is the *sine qua non* for AI adoption and, by extension, for the realisation of AI's societal benefits.

*" The European Union has fundamentally shifted the conversation. For the past years, we have been at the forefront of implementing secure and ethical AI frameworks while actively fostering the development of frontier technologies. At the EU, we support innovation, we support research, we support capacity building, and we support adoption. And above all, we are committed to striking the right balance between driving innovation and ensuring protection." — **Juha Heikkilä***

Then, Dr. Heikkilä traced the evolution of the EU's AI governance framework, beginning with the establishment of a high-level expert group on AI in 2016–2017. This group published ethical guidelines for trustworthy AI in 2019, which directly inspired the AI Act. The AI Act, he explained, is built on a risk-based approach: it bans certain uses of AI—such as social scoring and subliminal manipulation—that conflict with EU values, while imposing stringent requirements on high-risk applications. These requirements include data quality, traceability, transparency, human oversight, robustness, and security.

For general-purpose AI models (such as those underpinning chatbots and other widely used systems), the Act mandates transparency, ensuring that users are aware when they are interacting with AI. Systemic-risk models face even stricter obligations, including incident monitoring, risk management, and evaluations. To support compliance, the EU developed a Code of Practice in collaboration with over 1,000 stakeholders, led by experts including Yoshua Bengio. This Code, signed by 27 major companies, provides a practical pathway for providers to meet the AI Act's requirements.

On enforcement and international alignment, Dr. Heikkilä highlighted the EU's ongoing efforts to prepare for the AI Act's enforcement, including the selection of tenders for evaluating harmful manipulation and the establishment of an AI safety unit within the EU AI Office. He pushed back against the notion that AI safety has receded from the global agenda, pointing to the continued work of national AI safety institutes in Japan, Korea, Canada, and Australia, as well as the EU's own AI safety unit. The EU's approach, he noted, is to ensure that international initiatives—such as the G7's Hiroshima Process and the Global Digital Compact—align with its values and do not conflict with the AI Act.

The EU's international engagement is twofold: advocating for responsible AI governance globally and ensuring that external commitments do not undermine its own regulatory framework. Heikkilä acknowledged the complexity of the international AI governance landscape, suggesting that some consolidation of initiatives may be necessary to avoid duplication and ensure effectiveness.

In closing, Dr. Juha Heikkilä reaffirmed the EU's commitment to staying the course on the AI Act, despite calls for delays or omnibus revisions. He underscored that any adjustments are made for practical reasons—to facilitate compliance—not to dilute the Act's ambitions. The EU's approach, he concluded, is unique in its combination of ethical principles, hard law, and practical implementation, setting it apart from voluntary international commitments.

6. Ethical Foundations for the Governance of AI

Alexandre Mirlesse, Diplomat AI/Tech at the French Ministry of Foreign Affairs, opened the discussion by introducing Professor Alexei Grinbaum as a pivotal figure in AI ethics, describing him as a "Renaissance man" whose expertise spans quantum physics, robotics, music, philosophy, and multiple languages. Professor Alexei Grinbaum, Research Director at the CEA and Chair of its Digital Ethics Committee, also leads the AIOLIA project—a multinational consortium focused on translating ethical AI principles into practical norms and standards. The project's reach extends beyond Europe, including partners in Canada and elsewhere, reflecting a global ambition to operationalise AI ethics.

6.1 The Challenge of Translating Ethical Principles into Practice

Kick-starting the fireside chat, Alexandre Mirlesse began by reflecting on the earlier panel discussion, which had oscillated between the slow progress in establishing binding ethical standards and remaining optimism about incremental achievements. He asked Professor Grinbaum why it remains so difficult to produce ethical standards that genuinely influence the design and use of AI technologies.

Professor Grinbaum's response was direct: while there is no shortage of ethical frameworks—such as those from UNESCO, the EU, or national bodies—the critical missing step is operational translation. These frameworks often articulate high-level principles like transparency, explainability, and non-discrimination, but they rarely specify what these principles mean in concrete terms.

For example, what technical or organisational measures ensure transparency? How should discrimination be measured and mitigated? Professor Grinbaum argued that this gap between principle and practice is now being addressed, albeit gradually, through projects like AIOLIA, which aim to develop actionable guidelines for specific use cases—whether in medicine, education, security, or personal assistants.

He noted that this shift towards context-specific norms is already underway globally. China, for instance, has introduced regulations for virtual companions, while the EU is beginning to tailor its ethical guidelines to particular applications. The challenge, Professor Grinbaum explained, is to move from abstract values to portfolios of technical and organisational measures that engineers and designers can implement.

This process is complex, as it requires not only technical expertise but also an understanding of how ethical principles interact and sometimes conflict in real-world scenarios.

6.2 The "Heuristic of Fear" in AI Governance

Moving to the "heuristic of fear" in AI governance, Alexandre Mirlesse raised a provocative point from the panel: the idea that societal progress in AI safety might require a catastrophic event—a "Hindenburg moment"—to galvanise action. Professor Grinbaum rejected this fatalistic view, instead invoking the philosopher Hans Jonas and his concept of the "heuristic of fear". Jonas argued that societies can use the prospect of future catastrophes not as inevitabilities, but as motivators to prevent them. He extended this idea to AI, suggesting that the apocalyptic narratives surrounding the technology serve a political purpose: they create a sense of urgency that drives proactive governance.

"Why do we invoke apocalyptic visions when it comes to AI? Not to resign ourselves to their inevitability, but to make them credible enough to spur action. This is the political logic our societies apply to science and emerging technologies. We saw it twenty years ago with nanotechnology, and today, we are in the same spot with AI. These visions of catastrophe are not prophecies of doom; they are warnings that compel us to act now, to build the safeguards and governance that will prevent disaster. The goal is never to fulfil the apocalypse, but to render it impossible." — ***Alexei Grinbaum***

Going back to the tension between ethical principles and ethical AI governance frameworks, and to illustrate the complexities of operationalising ethics, Professor Grinbaum shared a hypothetical scenario involving a virtual companion. A young user, expressing distress to the AI, might say, *"I'm tired of you telling me to calm down. I don't want control—I just want you to listen."* Here, the AI faces a dilemma: respecting the user's autonomy (by not intervening) could conflict with a duty of care (by alerting a human supervisor if the user appears suicidal). He argued that such tensions cannot be resolved by abstract principles alone. Instead, they require contextual evaluation—assessing the user's intent, the severity of the risk, and the most appropriate response.

In this case, Professor Grinbaum suggested that LLM-based oversight might be more effective than human judgment, as AI systems can analyse subtle contextual cues that humans might miss. This example underscores the need for flexible, adaptive ethical frameworks that account for the nuances of real-world interactions.

6.3 The Need for Multidisciplinary Collaboration in AI Governance

Alexandre Mirlesse highlighted two key recommendations from Professor Grinbaum's recent publication: the necessity of collaborative, multidisciplinary processes and the involvement of the right people in ethical decision-making. He asked Pr. Grinbaum to elaborate: who are the "right people," and what skills or knowledge do they need?

Professor Grinbaum stressed that ethical AI governance requires institutionalised dialogue between at least two parties: the technical expert such as an engineer or designer and the ethics specialist who could be a philosopher, jurist, or policymaker. He cited the Ethics Readiness Levels (ERL) tool, developed for the European Commission, which helps project managers assess ethical risks by fostering dialogue between technical and ethical perspectives. For instance, an engineer might not immediately grasp how a system could influence human decision-making, but through structured discussion, these issues can be identified and addressed.

Professor Grinbaum praised France for establishing a National Digital Ethics Committee, a rare institution dedicated to facilitating such dialogues. He also noted the growing integration of ethics education in engineering schools and the inclusion of technical modules in humanities programmes, signalling a cultural shift towards interdisciplinary understanding.

To conclude, both Alexei Grinbaum and Alexandre Mirlesse argued that the pressing challenge is to cultivate a new generation of ethically literate engineers and technically informed humanists, capable of navigating between innovation and responsibility.

"As we approach pivotal political and diplomatic milestones—the 2027 French presidential election and the G7 Summit in June 2026—we bear a collective responsibility to elevate public debate in France on the ethical, economic, and scientific dimensions of artificial intelligence. Now, more than ever, conversations on AI are not optional, but essential. Only through open dialogue and inclusive decision-making can we ensure that AI serves the public good, reflects our shared values, and aligns with the aspirations of all citizens." — ***Alexandre Mirlesse***

7. Conclusion and Acknowledgment

The discussions held during the conference underscored the critical importance of ethical and responsible artificial intelligence. AI rapid development and deployment raise ethical, social, and political questions that demand our collective attention and decisive action. Now is the time to shape the future of AI in alignment with the common good. This conference served as a call to action for policymakers, industry leaders, and society at large—including academia—reminding us all that we have a role to play.

There is an urgent need to build governance models that bridge the gap between technical innovation and the broader societal impact of AI. The future of AI is not just about algorithms; it is about the choices we make for our societies, our economies, and our values. We must—and we can—harness the power of AI while upholding the principles of equity, transparency, and human dignity.

Against this backdrop, the speakers' interventions converged on the need for binding international safety standards, increased investment in AI Safety research, the inclusion of all nations in the global AI policymaking process, and the establishment of transparent, third-party audits for high-risk AI systems. The overarching message was clear: the time for voluntary pledges is passing; the path forward must be paved with concrete actions, shared responsibility, and a steadfast commitment to the public good.

"2026 must be the year of Global AI Governance in action, ensuring that rapidly advancing AI capabilities deliver a net positive impact worldwide. Bridging the gap between policymakers, researchers, practitioners, and civil society is critical to keeping AI safe, reliable, and beneficial for all." — ***Pauline Charazac***

The success of this conference, awarded the G7-label by the French Presidency, is a testament to the collective effort and dedication of all those who contributed to its organisation and execution. The organisers extend their deepest gratitude to each participant for their insightful contributions and unwavering commitment to advancing the cause of ethical and secure AI governance.

We are particularly indebted to our distinguished speakers, moderators and guests of honour whose expertise and vision enriched the dialogue and inspired all in attendance of the conference, under the patronage of Professor Natacha Valla, Dean of Sciences Po School of Management and Impact. Our heartfelt thanks go to Jessica Larsson, Chief of the European Commission Representation in France; Lord Tim Clement-Jones, Member of the UK House of Lords and Co-Chair of the UK AI Parliamentary Group; Eenam Gambhir, Deputy Chief of Mission of India to France; Miriam Minder, Co-Lead for Digital and New Technologies at the Swiss Federal Department of Foreign Affairs; Professor Yoshua Bengio, Co-Chair of the United Nations Scientific Panel on AI and Founder of MILA; Dr. Justin Vaïsse, Founder and Director General of the Paris Peace Forum; Professor Alexei Grinbaum, Chairperson of the CEA Digital Ethics Committee; Dr. Juha Heikkilä, Advisor on AI to the Director General of the European Commission AI Office; Charbel-Raphael Ségerie, Executive Director of CeSIA and Expert to the OECD AI Group; Nicolas Miailhe, Co-Founder of AI Safety Connect and Expert to the OECD AI Group; and to the Mardis de l'IA team, namely, Alexandre Mirlesse, Anne-Sophie Bordry, Louis Abraham et Théophile Cabannes as well as to the School of management and impact of Sciences Po, especially to Florent Bonnaventure, Carole Pizzinato, and Clara Amitrano.

Finally, we are grateful to the 135 registered participants—policymakers, diplomats, researchers, industry leaders, and students—whose active engagement and diverse perspectives made this G7-labeled conference a vibrant forum for exchange and collaboration. As we look ahead to the 2026 UN Global Dialogue and the 2027 AI Summit in Geneva, the insights and commitments generated here will serve as a foundation for continued progress as we remain committed to fostering international cooperation and invite all stakeholders to join in this essential endeavour.

8. Annexes

Detailed Agenda

– *Welcome at Sciences Po Innovation Pavilion & Family Picture with Speakers* **–**

17:00 – 17:10 | Welcome Address

Prof. Natacha Valla, Dean, Sciences Po School of Management and Impact

17:10 – 17:20 | Opening Remarks

Prof. Yoshua Bengio, Co-Chair of the United Nations Scientific Panel on AI, 2018 Turing Award and Founder of MILA
Introduced by **Charbel-Raphael Ségerie**, Executive Director, CeSIA & Expert to the OECD AI Group

17:20 – 18:30 | High-Level Roundtable (incl. 10min Q&A)

- **Lord Tim Clement-Jones,** Member, UK House of Lords and Co-Chair of the UK parliamentary group on AI
- **Dr. Justin Vaïsse**, Founder & Director General, Paris Peace Forum [Moderator]
- **Eenam Gambhir**, Deputy Chief of Mission of India to France
- **Miriam Minder,** Co-Lead, Digital and New Technologies, Swiss Federal Department of Foreign Affairs
- **Nicolas Miailhe**, Co-Founder, AI Safety Connect & Expert to the OECD AI Group

18:30 – 18:40 | Keynote Address - The European Union in Action: Turning AI Principles into Practice

Dr. Juha Heikkilä, Advisor on AI to the Director General of the European Commission AI Office

18:40 – 19:15 | Fireside Chat (incl. 10min Q&A)

Prof. Alexei Grinbaum, Chairperson, CEA Digital Ethics Committee
Moderated by **Alexandre Mirlesse**, Diplomat AI/Tech, Ministry of Foreign Affairs of France, AI Envoy (2026 Africa Forward Summit)
**The Fireside Chat was conducted in French. Q&A took place in both English or French.*

– *Closing and Thank you by Pauline Charazac, Head of Policy Engagement, CeSIA* **–**

Centre pour la Sécurité de l'IA

LABELLISÉ G7 FRANCE ÉVIAN 2026

SciencesPo SCHOOL OF MANAGEMENT AND IMPACT