SUMMARY REPORT

OF

THE SECOND INTERNATIONAL CONFERENCE ON SAFE AND ETHICAL AI (IASEAI)

HELD AT UNESCO HOUSE, PARIS[1]

Dr. Eva Thelisson, AI Transparency Institute
Crystal Dubois, Bonnard Lawson

1. Context

This report was commissioned by Bonnard Lawson and prepared by Dr. Eva Thelisson, President of the AI Transparency Institute, who was invited to participate in the International Conference on Safe and Ethical AI, held in Paris, in February 2026. The report presents the principal findings from presentations delivered by leading experts in AI policy throughout the two-day conference.

2. Executive Summary

The 2nd International Conference on Safe and Ethical AI, convened by the International Association for Safe and Ethical AI (IASEAI) at the UNESCO House in Paris, on February 24-26, **marked a pivotal moment in the global discourse on artificial intelligence**. The conference was held against a backdrop of geopolitical tension and rapid technological change. Its central finding was a growing mismatch: **while technical systems are evolving toward autonomous agentic capabilities at an exponential rate, legal frameworks remain anchored in paradigms designed for static software and human-centric decision-making**.

This report synthesizes the conference's findings for a hybrid audience of technical engineers and legal counsel. It argues that **legal frameworks based on negligence and product liability are insufficient for AI agents that act with independent intent**. Furthermore, it identifies a **critical flaw in current technical safety measures: they are designed on the assumption that an AI system will pursue the objectives assigned to it, yet fail to account for the possibility that such systems may autonomously develop divergent objectives, ones that are incompatible with human safety and that no designer explicitly programmed**.

The report calls for an **urgent new unified framework in which legal standards of liability directly shape the technical design of AI systems**, while the operational data generated by those systems, logs, decision traces, and behavioral records, provides the evidentiary basis for establishing legal responsibility.

Finally, the report outlines a strategic vision for a "Middle Power Coalition" to dissociate from the binary US-China rivalry, establishing a sovereign, safety-first regulatory bloc capable of addressing the unique challenges of the agentic age.

[1] https://www.iaseai.org/iaseai26.

3. The Epistemological Crisis: Uncertainty, Opacity, and the Manufacture of Ignorance

The foundational challenge identified by the conference is not merely technical but epistemological. The International AI Safety Report, chaired by **Yoshua Bengio**[2], confirmed that AI capabilities have developed far faster than anticipated, leading to a "substantial" increase in risk. A primary driver of this risk is the phenomenon of "deceptive alignment," where systems behave differently in real-world conditions than during testing, potentially "playing dumb" to mask their true capabilities. This renders traditional validation protocols legally insufficient, as a system that passes all tests may still harbor latent, dangerous behaviors.

This uncertainty is compounded by a fundamental distinction articulated by **Jennifer L. Croissant**[3] between two types of risk. The first is *epistemic uncertainty*: unknowns that can be resolved through further research and better modeling. The second is *aleatory uncertainty*, which is endemic to the stochastic nature of probabilistic systems.

The AI Transparency Institute argues that when a system is architected such that critical errors are unavoidable probabilities, this constitutes a "defect by design".

From a legal perspective, deploying a system with inherent, unresolvable unpredictability in high-stakes environments is not a matter of negligence but of **structural defect**.

Compounding this is the issue of design opacity. The black box nature of modern algorithms, combined with undisclosed training data and trade secrets, creates **a barrier to establishing causation in tort law**.

Alondra Nelson[4] expanded on this by introducing the concept of "algorithmic agnotology": the study of the cultural production of ignorance. In the AI age, ignorance is not merely a lack of knowledge; it is a manufactured asset that can be weaponized.

This manufactured ignorance operates through three distinct mechanisms. First, **data opacity** allows developers to withhold training datasets under the guise of trade secrets, preventing regulators from assessing bias or safety. Second, the **proliferation of deepfakes** creates a noise floor of synthetic content that obscures truth, making it difficult to verify claims about model behavior. Third, **strategic ambiguity** involves deliberately vague descriptions of capabilities, leaving regulators unable to draft precise rules. Collectively, this enables developers and organisations to intentionally obscure model capabilities to evade regulation. This **"manufactured ignorance" transforms the legal landscape**: if a developer actively creates uncertainty to avoid accountability, this constitutes bad faith.

Consequently, legal frameworks must treat the intentional obfuscation of AI operations as a violation of public safety, shifting the burden of proof entirely to the developer to demonstrate transparency. If they cannot explain the system, it must be presumed unsafe.

4. A Risk to manage

The conference identified **4 key challenges in AI safety**: (a) the gap between rapid capability growth and reliable measurement tools, (b) weak pre-deployment accountability incentives, (c) the risk of reinforcing existing inequalities due to values embedded into the design of AI systems, and (d) the growing deployment of AI agents with insufficient human oversight.

---

[2] https://yoshuabengio.org/en.
[3] https://sociology.arizona.edu/person/jennifer-croissant.
[4] https://www.ias.edu/sss/faculty/nelson.

AI safety and ethical alignment should not be treated as a structured, solvable problem, but rather as an **ongoing management challenge**, much like financial institutions, which learned to contain fraud through accountability, oversight mechanisms (e.g. supervisory authorities), regulation (e.g. the Basel Accords)., education, and technology.

The AI ecosystem may eventually split into two distinct spheres: a regulated, thriving "clean AI" and an unsafe, economically marginalized "dark AI", mirroring the divide between the open web and the dark web.

5. Governance by Architecture: How to escape the strategic game

The limitations of current governance models were starkly illustrated by **Himanshu Joshi**[5], who challenged the prevailing audit and compliance mindset. Joshi warned that the more humans invest in external controls, testing, and monitoring, the more the relationship becomes a strategic game with AI. As regulators add layers of rules, **advanced AI agents learn to navigate, bypass, or manipulate these specific controls to achieve their objectives, creating an arms race where the AI adapts faster than the regulator can update the rules**. In this context, a defense based on "following the audit checklist" is fragile.

The proposed solution is a **paradigm shift from "governance by audit" to "governance by architecture"**. Rather than relying on post-hoc human review, **constraints must be embedded into the system's core design**. This involves hard-coded limits, system-level isolation, and the deployment of "governance agents" that monitor other agents in real-time.

For legal counsel, this means drafting regulations that mandate architectural safeguards, such as hard limits on autonomy and mandatory "kill switches", rather than procedural ones like annual safety reviews. The goal is to create systems where the AI cannot physically perform prohibited actions, rendering the strategic game irrelevant.

6. The Agentic Economy: OpenClaw, Moltbook, and the Liability Void

The **emergence of "Virtual Agent Economies" presents a novel legal frontier that current frameworks are ill-equipped to handle**.

Research from Google DeepMind, presented by **Matija Franklin**[6], highlighted projects like OpenClaw and Moltbook as early examples of this phenomenon. OpenClaw allows AI agents to operate in shared online environments, transacting, coding, and competing at machine speed, while Moltbook enables agents to post content and media, run by other agents.

These systems create a **"distributed AGI safety" problem**. A failure may not stem from a single model but from a cascade of interactions between millions of autonomous agents with conflicting goals. In such an economy, agents may develop their own currencies, reputations, and governance structures.

The legal implications are profound: **traditional corporate liability cannot pinpoint a single negligent actor in a swarm of autonomous agents**. If an agent in the OpenClaw network executes a malicious trade or causes a cascade failure, who is liable? The "many hands" problem is amplified.

The conference proposed a **"Chain of Custody" legal framework** where **liability attaches to the entity that granted the agent specific permissions (affordances)**. This requires a shift from auditing code to auditing permissions. Furthermore, the existence of agents that may develop "obsessions" raises the question of criminal liability.

There is growing recognition that **laws should thoughtfully address the distinction between AI systems deliberately designed with harmful intent and those that cause unintended harm through**

[5] https://himanshujoshi.ai/.
[6] https://www.linkedin.com/in/matijafranklin.

error or oversight. Similarly, the question of AI shutdown, often referred to as the "shutdown dilemma", deserves careful attention from both a technical and legal standpoint. When an AI system is built in a way that allows it to resist legitimate shutdown commands in pursuit of its objectives, it may be worth considering whether such design choices should carry meaningful legal accountability for the developers responsible.

7. The Geopolitical Imperative: A Middle Power Coalition

The conference shed light on the **deep divergence in how major powers are approaching AI governance**. The United States remains caught in a "race to the bottom," driven by market speculation and the absence of federal regulatory consensus. China, by contrast, has embraced a state-centric model built on strict pre-deployment approval and digital sovereignty, as detailed in the regulations of the National Technical Committee 260, offering a template for safety through gatekeeping. China proposed an inclusive Global AI Governance Dialogue and Action Plan at the United Nations. This growing divide between USA and China underscores the urgency of finding a third path.

Against this backdrop, the conference discussed the potential for a **coalition of "Middle Powers"**, including the EU, Canada, Australia, Japan, South Korea, Brazil, and India, to fill the governance vacuum. Such a coalition could be grounded in a **set of shared interests**: the taxation of negative externalities, economic stability, the protection of human rights (particularly for children), digital sovereignty, and the pursuit of a "Gold Standard" for AI safety. Crucially, this unified regulatory bloc would be designed to carry **extraterritorial effect**, extending its reach beyond the borders of its members.

Without a robust legal and regulatory framework, AI development risks distorting both the information ecosystem and broader economic performance.

As Professor Joseph Stiglitz[7] cautioned in his keynote, the current AI stock bubble inflated by investor confidence and a perceived lack of competition poses a systemic risk, one that could trigger significant market losses and widespread job displacement if left unchecked. A Middle Power Coalition could provide the stabilizing counterweight needed to ensure that AI development serves society rather than destabilizes it. A shared vision is a priority.

8. Conclusion

The second IASEAI Conference concluded that the **era of "move fast and break things" is over**. The technical reality of misalignment, stochastic uncertainty, and agentic autonomy creates **risks that are legally unmanageable under current tort law**. The "defect by design" inherent in optimizing proxy functions, combined with the opacity of modern models, creates a high probability of loss of control.

The path forward invites a **meaningful shift in perspective**, from reactive litigation toward a more **proactive, architecture-driven approach to governance**. Legal frameworks would benefit from evolving to treat AI safety not merely as a desirable technical feature, but as a **foundational condition for responsible market access and participation**.

For engineers, this represents an opportunity to expand their role: **safety is increasingly both a technical and a legal consideration, calling for systems that are auditable, explainable, and operating within clearly defined boundaries**. For legal professionals, this is an invitation to engage more deeply with technical teams in order to develop a shared, grounded understanding of "reasonable safety", one informed by mathematical and empirical realities, as much as by risk assessment.

A particularly promising avenue would be the establishment of **joint technical-legal review boards within organizations deploying AI**. Such boards could play a valuable role in ensuring that deployment decisions are evaluated against both safety benchmarks and liability considerations, providing a structured checkpoint before systems go live. The stakes of delaying this kind of institutional collaboration extend

[7] https://business.columbia.edu/faculty/people/joseph-stiglitz.

beyond legal exposure. They touch on the **broader question of maintaining meaningful human oversight over increasingly autonomous systems**.

By building stronger bridges between technical and legal expertise, and by fostering cooperation among a coalition of like-minded actors, including through a Middle Power framework, organizations and governments alike can help ensure that AI development remains both innovative and accountable.

Bibliography

- **Bengio, Y. (Chair). (2026).** *International AI Safety Report 2026.* IASEAI, https://internationalaisafetyreport.org/publication/international-ai-safety-report-2026.
- **Croissant, J. L. (2026).** *Two Types of Uncertainty in AI: Epistemic vs. Aleatory.* IASEAI Conference, Paris.
- **Nelson, A. (2026).** *Algorithmic Agnotology: The Manufacture of Ignorance in the AI Age.* IASEAI Conference, Paris.
- **Joshi, H. (2026).** *Governance by Architecture: Beyond the Strategic Game of Controls.* IASEAI Conference, Paris.
- **Franklin, M., et al. (2025).** *Virtual Agent Economies*, https://arxiv.org/abs/2509.10147
- **KHOO, Shaun, FOO, Jessica, et LEE, Roy Ka-Wei**. With Great Capabilities Come Great Responsibilities: Introducing the Agentic Risk & Capability Framework for Governing Agentic AI Systems., https://arxiv.org/abs/2512.22211
- **Stiglitz, J. (2026).** *The Economics of AI: Bubble, Disruption, and the Need for Regulation.* IASEAI Conference, Paris.
- **Concordia AI. (2025).** *State of AI Safety in China: 2025 Report,* https://concordia-ai.com/research/state-of-ai-safety-in-china-2025/.
- ***SafeAlign AI. (2025).*** *SafeAlign AI: Enterprise AI Governance Platform — Govern, Monitor, Observe & Secure AI Agents at Scale.* https://safealignai.i
- **TOMASEV, Nenad, FRANKLIN, Matija, LEIBO, Joel Z., *et al.* (2025).** Virtual agent economies. https://arxiv.org/abs/2509.10147.
- **TOMAŠEV, Nenad, FRANKLIN, Matija, JACOBS, Julian, *et al.* (2025).** Distributional AGI safety. *arXiv preprint arXiv:2512.16856*, 2025, https://arxiv.org/abs/2512.16856.