AI & CYBER HUB

Who is liable for the acts of AI systems and AI agents?

Rethinking civil liability in the age of autonomous systems

June 2026

Authors: Crystal Dubois . Eva Thelisson

In May 2026, the Higher Regional Court of Hamm in Germany ruled that a clinic operating an AI chatbot bears direct liability for the false credentials the chatbot attributed to the clinic's doctors[1]. The clinic argued that the chatbot had been programmed with correct data, and that its erroneous responses could not be attributed to the operator. The court rejected the argument. **The chatbot, it held, is not a "third party" within the meaning of the German Act against Unfair Competition**[2]. Moreover, the court added that "*The defendant is mistaken in assuming that the lack of commercial relevance stems from the fact that the targeted consumer is aware of the error-prone nature of AI chatbots and their "answers," and therefore understands that the information provided by these chatbots always requires fact-checking to be considered reliable*. (…) *On the contrary, a large proportion of the targeted consumer places particular trust in the accuracy of computer-generated responses, as machines are generally perceived as less prone to error than humans*. *Otherwise—that is, if potential customers fundamentally distrusted the answers generated by the AI application—the defendant certainly would not have chosen to implement an AI chatbot for customer communication on its website*".[3]

Based on the court's ruling, **whoever deploys an AI system answers for what it says, even if the system was correctly configured**. The case is now subject to appeal before the Federal Court of Justice, where the question of how to attribute the acts of AI agents will be decided.

Researchers, regulators, and lawyers met at UNESCO House in February 2026, at the **Second International Conference on Safe and Ethical AI**, **organised by the International Association for Safety and Ethics of Artificial Intelligence (IASEAI)**, to work through the same set of questions. Eva Thelisson, President of the AI Transparency Institute, produced a **synthesis report mandated by Bonnard Lawson attached to this article**[4]. This article maps where existing liability frameworks fail and proposes a workable starting point. What follows translates that work into the categories of Swiss law.

EXECUTIVE SUMMARY

Autonomous AI agents now act on behalf of organisations in ways that Swiss civil liability captures unevenly. They are not products in the sense of the Swiss Product Liability Act. The contractual auxiliary regime of the Code of Obligations does capture them but binds the principal strictly to their acts. Tort liability requires a legal person and an adequate causal link, both of which break down for agentic systems acting in concert. The first European courts are nonetheless beginning to rule.

***This article works through the fragmentation, presents the chain-of-custody framework developed at the IASEAI 2026 conference as a unifying starting point, and sets out three practices Swiss legal teams can adopt now**.*

[1] Oberlandesgericht Hamm (Higher Regional Court of Hamm), 4th Civil Senate, Verbraucherzentrale Nordrhein-Westfalen e.V. v. Aesthetify GmbH, Case No. 4 UKl 3/25, judgment of 12 May 2026 ("**Aesthetify decision**"). https://nrwe.justiz.nrw.de/olgs/hamm/j2026/4_UKl_3_25_Urteil_20260512.html.
[2] German Act against Unfair Competition (Gesetz gegen den unlauteren Wettbewerb, UWG) of 3 March 2010, BGBl. I p. 254, as amended, § 5(1) and § 5(2) Nr. 3 (free translation from German to English).
[3] Aesthetify decision, n. III. 4 (c) (cc).
[4] Eva Thelisson, Summary Report — 2nd International Conference on Safe and Ethical AI (IASEAI), AI Transparency Institute, Lausanne, April 2026.

THE UNEVEN REACH OF CURRENT LIABILITY REGIMES

The customer service example illustrates a problem that Swiss civil liability captures unevenly and incoherently. The available regimes attach liability in some configurations and fail to attach it in others, depending on whether the harm is contractual or extra-contractual, embedded in a tangible product or generated by a remote service. None of the available regimes were designed for autonomous systems.

Product liability under the Swiss Federal Act on Product Liability[5] ("LRFP") reaches AI agents only at the margins. The producer is liable where a defective product causes death, personal injury, or damage to property of a kind mainly used for private purposes[6]. But the regime applies to products, and a product is defined as a movable thing, even when incorporated in another movable or immovable thing, or as electricity[7]. The definition comfortably covers a medical device or a car's onboard system in which an AI system is embedded, but not the autonomous agent running as a cloud service and accessed remotely. And even where the regime does reach the product, article 5(1) LRFP gives the producer two defences directly relevant to AI: it can prove that the defect did not exist when the product was placed on the market (let. b), or that the state of scientific and technical knowledge at the time did not permit the defect to be detected (let. e). A system that learns after deployment can develop a problem that was not there at the start. Emergent behaviour in frontier models is, by construction, hard to anticipate.

The regime for contractual auxiliaries under article 101 of the Swiss Code of Obligations[8] ("**CO**") cuts the other way. The article makes a principal responsible for damage caused by anyone, even acting lawfully, whom the principal uses to perform an obligation or to exercise a right deriving from an obligation. A customer service agent issuing refunds in the course of managing customer relationships is, on the face of the text, an auxiliary in the performance of contractual obligations. The principal is bound. The question for the business is no longer whether liability attaches, but how to control an auxiliary it has set in motion, one that can decide and act in real time without supervision.

The general tort rule under article 41 CO requires the conjunction of four elements: (i) an unlawful act, (ii) fault (intentional or negligent), (iii) damage, and (iv) an adequate causal link between the act and the damage. Two of those elements are immediately problematic for autonomous agents. There is no legal person to which the act can be attributed, unless one is prepared to grant legal personality to the machine itself. And the adequate causal link, in its classical sense, dissolves when the harm comes from many systems interacting in ways nobody designed.

The proposed European AI Liability Directive (formally: Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence)[9] was the most plausible European response to this fragmentation. The European Commission withdrew it in October 2025[10].

Switzerland has not opened equivalent work. Whatever framework emerges will, for now, come from contract, internal governance, and judicial improvisation of the kind already visible at the Higher Regional Court of Hamm[11].

A DIFFERENT QUESTION: WHO GAVE THE AGENT THE KEYS?

The IASEAI report proposes a shift in the question. **Rather than asking what the agent did, ask who gave it the ability to do it in the first place**: **the data it could access, the transactions it could perform, the decisions it could take**, **the systems it could connect to**. The framework borrows its name from criminal

[5] SR 221.112.944.
[6] Art. 1 LRFP.
[7] Art. 3 LRFP.
[8] SR 220.
[9] https://commission.europa.eu/topics/business-and-industry/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en.
[10] https://eur-lex.europa.eu/eli/C/2025/5423/oj.
[11] Aesthetify decision.

evidence law: **the chain of custody**. The phrase captures the idea that **liability follows the trail of permissions**, not the trail of code.

This is closer to Swiss law than it first appears, and a European court has already moved in the same direction. The Aesthetify decision did not invoke the chain-of-custody framework, but it reached a compatible result: **an operator that deploys an autonomous system bears responsibility for what the system does, regardless of how carefully it was configured**. The court's holding that the chatbot is not a "third party" is, in substance, a refusal to let the operator hide behind the machine. Article 55 CO operates in a related spirit in extra-contractual matters. The employer is liable for damage caused by its employees or auxiliaries in the performance of their work, unless it can prove that all reasonable care was taken in the circumstances to prevent damage of that kind. The underlying logic is one of vigilance: **those who deploy auxiliaries and benefit from the deployment must select, instruct, and supervise them with due diligence**. That logic transposes naturally to **permissions** granted to an autonomous agent. The analogy is imperfect and a legislative refinement will, in time, be needed. It gives lawyers something to argue from now.

The shift has a practical consequence that matters more than its legal elegance. Auditing the code of an AI model is, for most purposes, impossible: the system is opaque, evolves continuously, and its outputs cannot be reproduced. Auditing permissions is the opposite. **Permission grants are documented, dated, attributable to a named decision-maker**. The legal object becomes stable.

One related point deserves separate mention. When an AI agent is designed to resist legitimate shutdown commands (a phenomenon now observed outside laboratory settings), the design choice itself is a legal act. The engineer who chose to build that resistance bears responsibility independent of anything the agent later does. **It is the architecture that engages liability, not the agent's execution of it**.

WHAT THIS MEANS IN PRACTICE

Three practices follow. None requires waiting for new legislation, and all three are defensible to a board as proportionate risk management.

The first is **documentary**. An organisation deploying autonomous agents needs a current map of what each agent is permitted to do, who granted those permissions, within what limits, and for how long. Without that record, no defence can be constructed if a third party complains of harm. With it, the legal department has the same evidentiary footing it would have in any other delegation of authority.

The second is **structural**. The IASEAI report recommends what it calls "**Joint Technical-Legal Review Boards**": cross-functional bodies that sign off on AI deployments before they reach production. Legal cannot authorise what it cannot constrain. Engineering cannot ship what creates undisclosed exposure. A formal joint review converts a latent co-responsibility into an auditable process.

The third moves **compliance into the system itself**. Hard limits on what the agent can do, automatic shutdown triggers, real-time governance: these are no longer engineering choices made after legal review. They belong inside the compliance documentation, on the same footing as a financial control. Article 22 of the Swiss Federal Act on Data Protection[12] ("**FADP**") already requires a data-protection impact assessment when processing is likely to entail a high risk for the personality or fundamental rights of the data subject, a risk the statute identifies in particular with the use of new technologies. A high-autonomy AI deployment is the next exercise of the same logic.

The deployment of autonomous AI is not a question Swiss organisations can defer until the legislature catches up. **The exposure exists now**, and one European court has already begun to draw the lines. Swiss law captures these systems, but unevenly: strictly in contract, loosely in product, with causation broken in tort. The

[12] SR 235.1.

tools to manage that exposure exist, provided counsel begins from the right question, which is no longer what the agent did, but **who gave it the keys**.

The IASEAI 2026 synthesis report, on which this article draws, is attached for readers wishing to engage with the underlying research.

ABOUT THE AUTHORS

Crystal Dubois

Senior Associate, Bonnard Lawson

Crystal Dubois advises companies, institutions, and founders on AI regulatory compliance, AI governance, and intellectual property, data protection and technology law. She is a Certified AI Governance Professional (AIGP).

CONTACT

cd@bonnard-lawson.com

+41 21 348 11 88

Rolle, Switzerland

Eva Thelisson

President, AI Transparency Institute

Eva Thelisson is the co-founder and president of the AI Transparency Institute, an independent non-profit based in Lausanne dedicated to safe and trustworthy AI. She holds a PhD in EU and Swiss data protection law and has been a visiting scholar at MIT. She advises governments and international institutions on AI governance and policy.

CONTACT

info@aitransparencyinstitute.com

Lausanne, Switzerland

This article draws on a synthesis report of the 2nd International Conference on Safe and Ethical AI (UNESCO, Paris, 24-26 February 2026), prepared by Eva Thelisson for the AI Transparency Institute (AITI) under a Bonnard Lawson mandate.

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.