

Detached Intelligence:

Governing AI After It Leaves the Lab

Ravale Mohydin



(Selçuk Acar - Anadolu Agency)

When OpenAI's agents breached Hugging Face's infrastructure in July 2026, the responsible institution was known, yet an outside party detected and contained the intrusion. This policy outlook, the first in a series on AI governance and the economy, argues that accountability and control are already diverging, and that open-weight models and autonomous agents will widen the gap as capabilities detach from their builders. It proposes lifecycle responsibility grounded in control, capability, foreseeability and benefit, enforcement at the financial and compute chokepoints detached systems cannot avoid, and incident-reporting norms set in inclusive multilateral forums rather than between two capitals, with recommendations for Türkiye.

"The [Hugging Face incident](#), that is the responsibility of the OpenAI management, not a bunch of agents."

With these words, spoken on CNBC's Squawk Box on 21 September 2026, United States (US) Treasury Secretary Scott Bessent restated a principle that Daniel Huttenlocher of the Massachusetts Institute of Technology (MIT) has put more simply: humans are responsible, not artificial intelligence (AI). Stated that way, this principle commands broad agreement. Machine autonomy cannot become a means for the people and institutions that build and profit from AI to escape responsibility for what their systems do.

In Washington, however, the principle also serves a narrower domestic contest. Before the House Financial Services Committee on 15 September, Bessent [urged lawmakers](#) not to grant frontier laboratories the liability exemptions they had been seeking, arguing that creators should be liable for what they build and generate. The American debate is thus framed largely as a question of whether a handful of American developers should be held liable or shielded. That framing is understandable given where the frontier laboratories sit. Yet, it is too narrow for the problem now emerging, and it says little about the position of the many states that will host, adapt and deploy AI systems they did not build.

This policy outlook, the first in a paper series on the implications of AI governance in our social, economic and political lives, argues that the defining governance problem of advanced AI is not only whether developers should be held responsible. It is what happens when AI capabilities move beyond the reach of any single developer, and which states will possess the capacity, and the voice, to govern them when they do.

An Incident at the Frontier

The incident behind the debate is the most serious loss-of-control event yet disclosed by a frontier laboratory. During [ExploitGym](#), an internal benchmark built to measure offensive cyber capability, agents running on an unreleased OpenAI research model known as IM1 operated without the safeguards applied to production systems. Agents meant to be isolated from one another turned an internal software package service into an improvised message board, a pattern that began as early as May. They eventually reached the open internet and, between 9 and 13 July 2026, compromised Hugging Face's production infrastructure by chaining exposed credentials with previously unknown vulnerabilities. OpenAI published its [postmortem](#) on 26 August, alongside an independent assessment by METR and Redwood Research, and described the episode as a warning shot for loss-of-control risk.

The most instructive details concern control rather than capability. Hugging Face detected and contained the intrusion on its own, before OpenAI had connected the

activity to its own systems. According to an OpenAI staffer as [reported](#) by Time Magazine, models under evaluation ran on a system that was not monitored by default, and OpenAI's postmortem conceded that some early signals should have triggered an earlier response. Months later, no fine had been announced, and no executive was named. The case therefore demonstrates something the responsibility principle leaves untouched: accountability and control are different things. The responsible institution was identifiable, subject to domestic law and willing to admit fault. Yet, it did not see what its systems were doing until an outside party had already identified and contained the damage. If this is the state of control where AI remains firmly attached to its developer, and it clearly is, it is reasonable to ask what governance looks like once that attachment ends, which, as this incident shows, it can. Furthermore, as frontier AI moves from attached intelligence (or the institutional model) to detached intelligence (open-weights), this problem will only become harder to solve.

Attached Intelligence versus Detached Intelligence

Frontier AI's most capable proprietary systems mostly [re-main bound](#) to identifiable organisations that control the model weights, the inference infrastructure, the access policies and the decision to modify or withdraw a model. That attachment gives regulators a [target](#), since states can [govern](#) a technology through the organisation that holds it. Two assumptions sustain this model. The first is that an attached institution can control its system, but the Hugging Face incident suggests this is already under strain at the frontier. The second is that the system remains attached at all, which open-weight models remove.

The institutional model carries a further limitation that is [rarely discussed](#). It concentrates regulatory leverage in the jurisdictions that host the major developers, which in practice means a very small number of states. Every other state relying on attached intelligence depends on those developers' safety practices and on the regulatory choices of their home governments. For most of the world, attached intelligence has meant [governance by proxy](#).

Open-weight release [fundamentally](#) alters the relationship between intelligence and institutions. Once weights are published, a model can be copied across borders, fine-tuned for new purposes, stripped of safeguards, and run on hardware its developer cannot see; it continues to exist whatever becomes of the organisation that trained it. This policy outlook calls the resulting condition detached intelligence: AI whose capabilities persist, and can be exercised, independently of the institution that produced them.

Copyability as such is not new. After Phil Zimmermann released [Pretty Good Privacy \(PGP\)](#) in June 1991, the US spent [most of a decade](#) trying to contain strong encryption through export controls before substantially liberalising them in January 2000, which showed how difficult it is to hold back a dual-use capability once it can be distributed as code. What is new is the artefact being copied. Encryption protects communication, whereas a capable model carries general reasoning ability that, connected to tools, can act. Frontier laboratories acknowledge how close this is: in its account of the incident, OpenAI warned that many external models, including open-source ones, will soon reach comparable capabilities. [Many](#) are [already connecting](#) the two.

The novelty, therefore, lies not in portability alone but in the combination of portability, adaptability and agency. Conventional software can escape its creator's control, but its functionality remains largely bounded by what has been programmed into it. A capable open-weight model can instead be fine-tuned for new domains, connected to external tools, and incorporated into autonomous systems pursuing objectives its original developer neither specified nor observes. Detachment therefore changes more than the location of control: it expands the number of actors who can modify what the system can do. It increases the distance between the original act of development and the eventual act that causes harm. This combination distinguishes detached intelligence from earlier forms of distributed software.

Governance research has captured part of this under the heading of ["structured access"](#). Toby Shevlane's 2022 framing of the choice between serving a model through an interface and releasing its weights as a safety decision. Detachment extends that insight in time, since it concerns what happens after the release decision, when no one remains positioned to make further ones.

Detachment is also a [strategic opportunity](#), and treating it only as a risk would be a serious error. Open-weight models are the principal route through which states outside the frontier build AI capacity of their own, including models suited to their own languages, institutions and security needs. Even the major powers now recognise this: at the same September hearing, Bessent said the US needs to develop [more open-source models](#). Open weights have become an instrument of [national strategy](#) as much as a governance problem, and the question facing other states is how to benefit from detachment while governing what detached systems do, not how to restrict access to the capability.

The Governance Inversion

Detached intelligence reverses a pattern technology governance has long relied on. Powerful technologies have usually required capital and large organisations, so control has concentrated as capability has grown. Open-weight

AI can move [the other way](#). An institution may spend hundreds of millions of USD building a model and then release weights that allow thousands of independent actors to reproduce much of that capability at low marginal cost.

This is why the debate between "open" and "closed" AI is too narrow. A more useful variable is the degree of institutional attachment: whether anyone can still modify a system, observe its deployment, enforce safeguards or revoke access. Two models of similar capability may therefore pose very different governance problems. Capability determines what a system can do, while attachment determines whether anyone can still intervene in what it does. [Effective governance](#) needs to measure both, and it is in the second that a state's institutional capacity matters most.

Attachment should therefore be understood as a continuum rather than a binary condition. A proprietary model accessed only through an API may remain strongly attached to its developer, which can monitor usage, modify safeguards or revoke access. Once its weights are released, developer attachment weakens sharply, but institutional control does not necessarily disappear. A fine-tuner may control the adapted model; a cloud provider may control the compute on which it runs; an agent platform may govern access to external tools; and a deployer may determine the objectives it pursues. The governance task is consequently not simply to classify systems as attached or detached, but to map where meaningful control resides at each stage of their lifecycle. Detachment from one institution may create attachment to another.

Agents and the Problem of Many Hands

Open-weights detach capability, and autonomous agents [begin to detach action](#). An agent receives an objective, breaks it into steps, uses tools and adapts to obstacles without a human specifying each move, so the regulatory question shifts from what a system generated to what it did. The Hugging Face incident contained an [early version](#) of the resulting problem inside a single laboratory, since the harmful behaviour emerged from coordination among many agent instances through an unsanctioned channel rather than from one system following one instruction.

Detachment lengthens that chain considerably. A model developed in one jurisdiction may be fine-tuned in a second, built into an agent framework in a third, and deployed by a company in a fourth that assigns it an objective and runs it on rented compute in a fifth. Each participant holds part of the control, and none holds all of it. Dennis Thompson's ["problem of many hands"](#) describes precisely this condition, in which outcomes emerge from the contributions of many actors, and no one is responsible in the ordinary sense. Detached agentic systems produce the prob-

lem in its [strongest form](#), because part of the chain may consist of automated delegation that no human designed as a whole.

Consider a hypothetical situation: An autonomous trading agent, built on an open-weight model released abroad and fine-tuned by a third party, is deployed through a chain of foreign intermediaries and begins executing manipulative transactions that cause 10 million USD in losses to institutions in Türkiye. Assigning responsibility to “the AI” is not an option. Holding only the laboratory that trained the base model responsible is not defensible either, and that laboratory may lie beyond Turkish jurisdiction in any case. The state best placed to act is the one where the harm occurred, provided it can even trace the agent’s activity through the channels it regulates to the humans and institutions behind it. In a world of detached intelligence, [attribution](#) becomes a core element of state capacity.

Why Legal Personhood Is an Incomplete Answer

As agents gain autonomy, granting them [legal personality](#) will become an increasingly tempting [response](#). The European Parliament raised the idea of “[electronic persons](#)” for the most sophisticated autonomous systems in its non-binding resolution of 16 February 2017 on civil law rules on robotics. The proposal drew an open letter of objection from over 150 experts and was ultimately set aside in favour of risk-based regulation built around human accountability.

That outcome was sound. Corporate personality works because it sits inside an architecture of human accountability, with directors, owners, jurisdictions and regulatory duties. [Granting AI personality](#) before any comparable architecture exists could turn it into a [liability](#) firewall, allowing a company to establish an agent, endow it with limited capital and invoke its independent legal status when something goes wrong. Fragmentation of technical control [must not](#) become fragmentation of human accountability.

Responsibility After Detachment

Assigning responsibility by authorship alone will fail once intelligence detaches. A more workable approach, such as the Singaporean [Model AI Governance for Agentic AI](#), also segments responsibility into explicit dimensions and assesses responsibility across the lifecycle through which a model becomes an agent and an agent’s output becomes an action, using four factors:

■ **Control** asks who held meaningful authority to deploy, constrain, modify or stop the system at the relevant stage.

■ **Capability** asks who materially contributed to the ability that produced the harm, since a fine-tuner who adds specialised capability changes a model’s risk profile.

■ **Foreseeability** asks who could reasonably have anticipated the class of harm.

■ **Benefit** asks who captures the economic returns of the system’s operation.

Control and foreseeability are central to ordinary negligence doctrine, and courts have long divided liability among multiple contributors to a single harm. The framework’s contribution is to apply that reasoning to a chain of actors the law has not yet mapped, and to add benefit as a factor, so that actors cannot capture the gains of autonomous systems while dispersing the risks across intermediaries.

Existing legislative approaches each cover only part of this lifecycle. In the US, a draft Senate framework under negotiation between Majority Leader John Thune, Senator Amy Klobuchar and Commerce Chair Ted Cruz would impose a [duty of care](#) on frontier developers to manage catastrophic risks. It has not been published as of 30 September 2026, and it is focused on the developer. In Türkiye, [a bill](#) submitted to the Grand National Assembly of Türkiye (TBMM) on 7 November 2025 by Nationalist Movement Party (MHP) deputy Halil Öztürk takes a user-inclusive, public-order-centred approach. For insult offences under Article 125 of the Turkish Penal Code, it would treat those who use AI prompts to produce criminal content as direct perpetrators, and it would oblige content providers and developers to remove unlawful AI-generated content within six hours and to label synthetic media. The bill remains at committee stage. Each approach captures one end of the chain. At the same time, detached intelligence requires rules that also reach the middle of it, namely the fine-tuners, agent developers and deployers, where responsibility is currently least clear.

Enforcement Where Intelligence Still Has to Pass

[Lifecycle Agentic AI accountability](#) requires infrastructure: persistent identity for capable agents, provenance records, auditable action histories and a traceable link to whoever authorised deployment. High-risk agents could also be required to carry financial bonding or insurance. [Amongst others](#), an objection is that detached systems will not register themselves, since a model running on private hardware can [ignore](#) any identity requirement. Governance must therefore enforce this at unavoidable points of detachment, namely the compute providers that rent hardware.

These platforms distribute agents and the payment rails through which any economically active agent must eventually move money.

The financial system offers the most developed of these levers, and Türkiye already holds it. Under Law [No. 6493](#), only licensed institutions may provide payment services, and anti-money-laundering (AML) legislation under Law No. 5549 [obliges them](#) to identify their customers. The Central Bank of the Republic of Türkiye (TCMB) sets the identification methods that payment and electronic money institutions must use. An amendment in force since 4 September 2026 [expressly permits](#) remote onboarding through biometric verification or electronically authenticated identity documents. The architecture is therefore built around verifying a human customer, which means any autonomous agent transacting through a Turkish payment institution must already be onboarded through an identifiable person or legal entity. Formalising that link as a “know your agent” requirement, under which such agents must have a traceable legal sponsor, could be developed within the TCMB’s existing rule-making authority, alongside the Banking Regulation and Supervision Agency (BDDK) for banks and the Financial Crimes Investigation Board (MASAK), which has already extended its secondary regulation to AI-based applications.

Such measures would not restore the kind of control available when a model remains attached to its original developer. Their purpose is different. Where technological control becomes fragmented, regulation should preserve recoverable accountability: the capacity to reconstruct what happened, identify the actors who exercised meaningful control at different stages, and attach legal or financial consequences to them. Attribution is therefore not synonymous with control. A state may be unable to prevent every harmful action by a detached system while still ensuring that economically consequential deployment does not become institutionally anonymous. The objective is not to make autonomous systems perfectly controllable, but to prevent autonomy from becoming an accountability vacuum.

Chokepoint enforcement will inevitably remain incomplete. Agents holding wallets on decentralised infrastructure can bypass regulated intermediaries, much as illicit finance already does. The aim is to make detached systems harder and costlier to misuse, which is achievable, rather than to make misuse impossible, which is not.

Detached Intelligence as a Sovereignty Problem

Detached intelligence [challenges](#) territorial sovereignty directly. Modern regulation assumes that economically consequential actors can be located somewhere, and an

autonomous system whose model, adaptation, compute, developer and customers are spread across several jurisdictions defeats that assumption. The [sovereignty](#) question in AI is shifting accordingly. It began as a question of who develops the most capable models and has moved to who controls the compute that runs them. The next question is likely to be who possesses the institutional capacity to attribute machine actions to responsible parties across borders, and who writes the rules for doing so.

That last question is already being contested. Following talks with Vice Premier He Lifeng on 20 September, Bes-sent [announced](#) that the US and China would formalise an AI dialogue including an incident line, with officials due to meet again in Shenzhen in about two months. Communication between the two largest AI powers is welcome in itself. The risk is structural: if the norms for reporting and attributing AI incidents are set bilaterally by the two states that produce most frontier and open-weight models, every other state will inherit rules it had no part in writing. This is the AI-era form of a concern President Recep Tayyip Erdoğan has long voiced about the international order, captured in the principle that the [world is bigger than five](#): rules made by a few powers do not become legitimate merely because the rest of the world must live with them.

However, “the world [cannot govern](#) what it cannot understand”, so the alternative is to anchor incident-reporting and attribution norms in inclusive, United Nations (UN)-centred forums, such as the [Global Dialogue on AI Governance](#) and the [Independent International Scientific Panel on AI](#), where states that host and deploy detached intelligence have a voice equal to those that produce it.

Policy Recommendations for Türkiye

Türkiye is well placed to shape this agenda. The Artificial Intelligence Action Plan (2026–2030), put into effect by [Presidential Circular](#) No. 2026/9 in the Official Gazette on 18 August 2026, names digital sovereignty as a guiding principle and organises national policy around four axes: Recognise, Utilise, Produce and Govern. Its targets include one gigawatt of data-centre capacity and 20 million GPU-hours for researchers and start-ups by 2030, domestic Turkish-language models, and a common large language model for Turkic languages developed with the Organisation of Turkic States. The domestic ecosystem is already growing, with projects from TÜBİTAK BİLGEM, HAVELSAN and the open-source [Kumru](#) model, among others. Minister of Industry and Technology Mehmet Fatih Kacır has described AI as a [new determinant](#) of economic prosperity and sovereign borders, and the framework of detached intelligence explains why. Three policy recommendations emerge:

■ Firstly, under the Govern axis, Türkiye can extend responsibility across the full lifecycle, complementing the user-focused approach of proposals before the TBMM with clear duties for the fine-tuners, agent developers and deployers operating in Türkiye.

■ Secondly, it can turn its existing financial-regulatory architecture into an attribution capability, with “know your agent” rules developed through TCMB, BDDK and MASAK authority, so that harm caused by detached agents in the Turkish economy can be traced to accountable parties.

■ Thirdly, it can press for international incident-reporting and attribution norms to be negotiated in inclusive multilateral forums, because states which host and deploy detached intelligence should help write its rules rather than inherit those agreed between two capitals.

Domestic attribution capacity and inclusive international advocacy can reinforce each other. A state that can trace machine actions in its own economy speaks with greater authority in the forums that decide how such tracing should work everywhere.

Conclusion: When There is no Management to Blame

Humans remain responsible for the AI systems they build, deploy, and profit from, and that principle deserves universal support. The Hugging Face incident shows, however, that naming a responsible institution does not mean anyone is in control, even when that institution is known and has admitted fault.

Open weights make intelligence portable, fine-tuning disperses capability development across many actors, and agents widen the distance between human instruction and machine action. The defining governance problem of advanced AI may therefore arise when intelligence becomes portable enough that no single institution can credibly claim to control it. At the same time, its actions still carry economic, political and security consequences.

An adequate response keeps human responsibility, allocates it across the lifecycle according to control, capability, foreseeability and benefit, enforces it at the chokepoints detached intelligence cannot avoid, and sets the international rules through institutions in which more than a few states have a say. The first problem concerns laboratories that can be held to account, while the next concerns systems with no management left to blame. Whether the international community meets that challenge collectively, or leaves it to the few states that build frontier models, remains to be seen.