“To Bot or Not to Bot”

Conscious Agentic Deployment in a Rapidly Changing World

Katryna Dow
July 2026

meeco

“When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash; and when you invent electricity, you invent electrocution. Every technology carries its own negativity, which is invented at the same time as technical progress.”

- Paul Virilio

Acknowledgement

Developed from a keynote delivered by Katryna Dow at KuppingerCole's European Identity and Cloud Conference (EIC 2026), Berlin, May 2026.

Contents

Foreword

In May 2026 I gave a keynote at the European Identity Cloud conference (EIC2026) in Berlin, on a stage KuppingerCole has hosted for nineteen years. The title of the keynote was the same as this paper. So, I want to start by acknowledging the opportunity to create and deliver that presentation, without that context none of what follows makes as much sense.

Martin Kuppinger, KuppingerColes's founder and Distinguished Analyst opened the conference this year with a line I haven't stopped thinking about:

"everything we didn't solve backfires with force now"

Sitting in the audience, I felt the full weight of his comment which inspired me to build on my keynote and write this paper.

Two decades of patient, unglamorous work on identity, federation standards, credential formats, trust frameworks, the endless working group calls about who governs what has been building toward exactly this moment, and most of the world doesn't know it yet. For those of us part of those communities we can see what is coming, everyone else is about to find out the hard way.

Here's what I mean. For thirty years, the identity community has been solving a human problem: how do you let a person prove who they are, once, and be trusted everywhere, without handing over more of themselves than the moment requires?

Whilst that work is not finished, it has matured and significant progress has been made. We have standards. We have wallets. We have regulations guiding governments to issue verifiable credentials at national scale. And just as that maturity was arriving, the problem changed shape entirely.

It is no longer only **who is the person**. It is **who is the agent acting on the person's behalf, and who answers for what it does.**

That is not a new committee item. It is a different problem wearing the identity industry's clothes, and it will not be solved without a conscious process before deployment.

Unfortunately, we are watching real-time as human-credential-sharing has fast become the default onboarding pattern for agentic commerce in the space of about eighteen months. It is easy to understand why… it's fast, it's what the tooling makes easy, and nobody budgeted for an entirely new alternative workforce. **But it is a liability structure with no name yet, sitting quietly inside almost every organisation that moved fast on agent deployment.**

So, this paper aims to do two things; firstly, to highlight why the identity discipline must be consciously applied to Agentic deployment. Secondly, to widen the lens

beyond the enterprise, because, more importantly, I don't think this is only a governance question given a technology this consequential moving this quickly.

Instead, it's a question about what we owe each other from a societal perspective when viewed through our roles as technologists, employers, regulators, patients, students, customers, passengers, parents and friends.

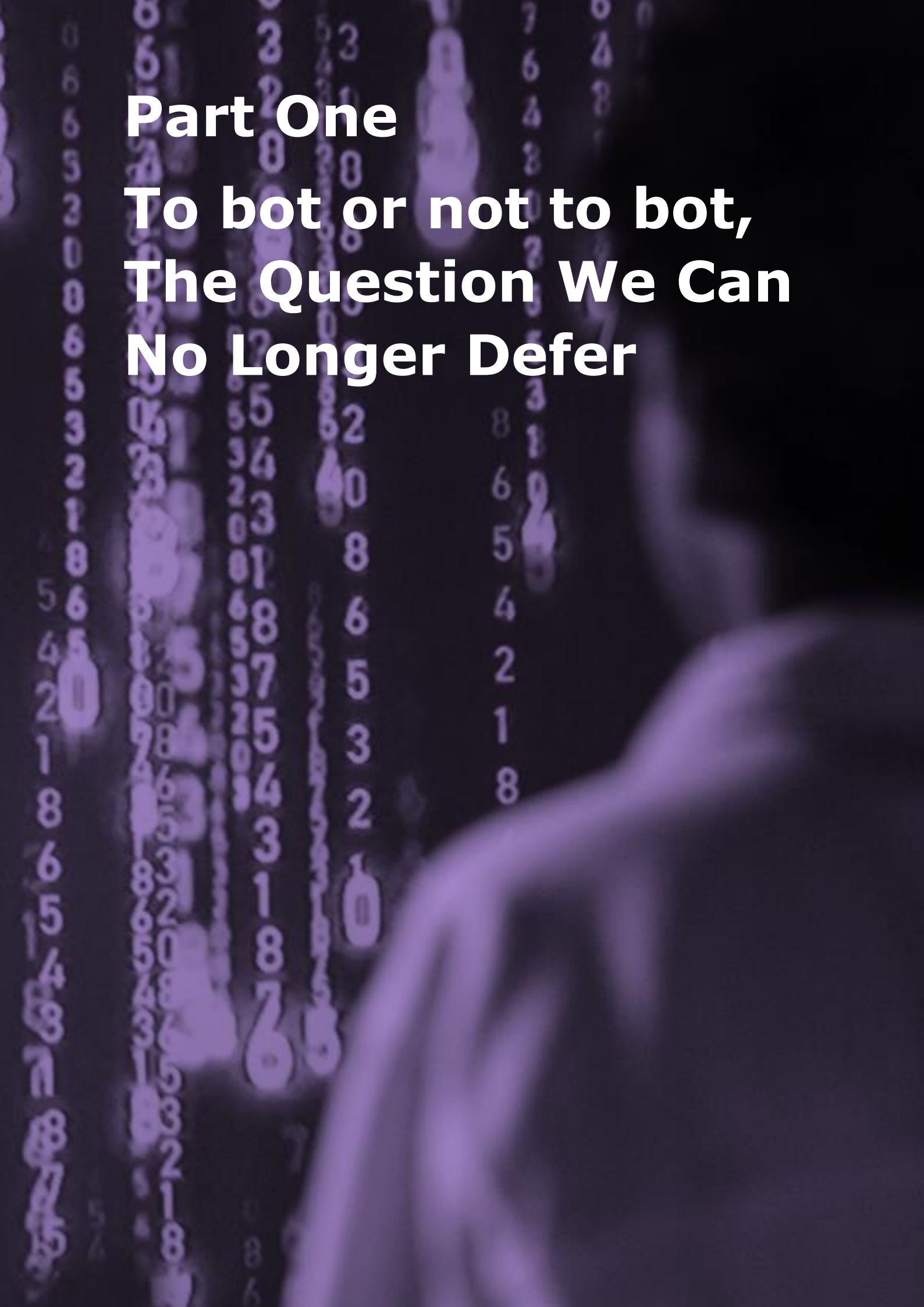
In sharing my perspective, my aim is to hold both the innovation case and the accountability case in the same hand, because I think anyone who only argues one side of it isn't being realistic. There’s no putting the AI genie back in the bottle, but at the same time, it is still early enough to mitigate some of the looming risks.

I don't have a tidy answer on striking that balance; however, I would like to propose a simple framework, five questions every organisation should be able to answer before an agent gets access to anything that matters.

I have a strong suspicion that the answer to "to bot or not to bot" is going to be **maybe** for a good while yet, which when compared to either the boosters or the doomsayers, makes **maybe is a good place to start**.

Once again, thank you to KuppingerCole and the EIC community, and the two decades of groundwork this paper stands on.

Katryna Dow
CEO & Founder, Meeco
Brussels, July 2026



Part One

To bot or not to bot, The Question We Can No Longer Defer

Part One: To bot or not to bot, The Question We Can No Longer Defer

“To bot or not to bot?” is possibly the Shakespearean question of our time. Shakespeare's Hamlet wrestled with action versus inaction and the weight of consequence. We are facing a version of the same question now, except the prince has been replaced by an algorithm, and the stakes are distributed across every organisation that has decided, in the space of a single product cycle, to let software act on its behalf.

Virilio's warning, is the right starting point, but it is worth being precise about where it runs out. Virilio's claim is that every technology invents its own negativity alongside its progress, the ship invents the shipwreck, the plane invents the crash. That is true, and it has been true of every technology humans have built. **But it describes a world in which humans still decide**. The ship doesn't choose to sink. The plane doesn't choose to crash. A human pilots it, badly or well, and bears the consequence either way.

> **“When you invent the ship, you also invent the shipwreck; when you invent the plane you also invent the plane crash; and when you invent electricity, you invent electrocution. Every technology carries its own negativity, which is invented at the same time as technical progress.”**
>
> **- Paul Virilio**

In his recent book, ***Nexus***, the historian Yuval Noah Harari, makes the argument that AI breaks this pattern for the first time in the historical record. His formulation is worth stating exactly; ***"AI can process information by itself and thereby replace humans in decision making. AI isn't a tool, it's an agent."*** The printing press could copy the Bible; it could not write a new one. The atom bomb was, in Harari's words, still a bomb, heads of state and their advisers decided where it fell, not the bomb itself. AI is different in kind, not just in degree: it can decide, and it can generate ideas no human proposed, without a human in the decision at all.

That distinction is the reason I wanted to write this paper. If AI were simply a faster, more capable tool, the governance conversation would be a scaled-up version of the one the identity industry has been having for years: better access control, better audit logs, better authentication. Important, but not new in kind. But if AI is, as Harari argues, an agent rather than a tool, then the question ***"who is the agent, and who answers for what it does"*** is not an implementation detail.

It is the **first** question, because for the first time, the answer might genuinely be "the software decided this, not a person", and if true, may have catastrophic

consequences. Consequences which may be mitigated if we take the time and apply a more conscious deployment approach.

It is worth being precise about how far "acting" now extends. AI agents are no longer confined to answering questions. They are taking actions, making decisions, spending money, sending emails, and managing access, often within workflows a human never directly reviews. Agentic Commerce, in the language now used across the industry, is here.

The pace of this shift has no real precedent. Compared to previous foundation technologies; mobile, personal computers, the internet, the rate of AI adoption is unparalleled. Gartner forecasts that 40% of enterprise applications will carry task-specific AI agents by the end of 2026, up from less than 5% in 2025, an eightfold jump in a single year, and separately projects that by 2028, roughly a third of user experiences will have shifted from native applications to agentic front ends entirely.

Salesforce's own platform data tells the same story from the demand side: agent creation among early-adopter businesses grew 119% in the first half of 2025 alone, and Australian consumers who regularly use AI agents report 64% higher satisfaction than those who don't, a sign of genuine, fast-forming consumer readiness, not just enterprise enthusiasm. This is not a slow-burning technology adoption curve. **It is closer to a phase change**.

That creates a problem that has nothing to do with whether the technology works. It works, mostly. The problem is that an agent without an identity is an actor without accountability, and that is not a technical footnote, it is a problem that reaches into how organisations assign liability, how regulators write law, and eventually how society decides what it is comfortable delegating to something that cannot be held to account the way a person can.

So, this rapid shift is not really just about technology, nor are the challenges just tech problems. This size of shift in such a short period of time is something that's really impacting the whole of society.

Before any organisation deploys an agent to act on its behalf, something that takes, in practice, seconds, it is worth pausing on a genuinely uncomfortable question: **do we understand what we are unleashing?** Not rhetorically. Literally: if this agent does something we didn't intend, can we trace it, stop it, and explain it to a regulator, a customer, or a court?

Are we taking a pause to think about the consequences and the downstream actions that may happen? Yes…no…maybe? **My aim is to make it easier to apply practical steps in order to shift intentionality further upstream.**

AI agents have moved from answering questions to taking actions: booking, transacting, spending, and accessing systems on behalf of real people and organisations. That shift has outrun the infrastructure meant to govern it.

Most agent deployments today still rely on a decades-old workaround, sharing a

human's credentials with the agent, which means that when something goes wrong, there is often no reliable way to say who or what was responsible.

On the following pages I have traced that problem across three connected fronts; identity, governance and the bigger societal picture. This is followed up with a practical framework that I hope will help to close the gap.

None of these are side stories, this is the terrain any enterprise AI strategy must now navigate.

Part Two

Who is the Agent: The Identity Problem

Part Two: Who is the Agent: The Identity Problem

An agent, not a tool

Every enterprise identity system in production today was built on the assumption that the entity on the other end of an authentication flow is a human, or an algorithmic piece of infrastructure that behaves like one. That assumption has held for years because it was true. However, it stops being true the moment the "user" on the other end of the session is something that can decide, adapt, and act without a human approving each step.

This is where Harari's distinction stops being a philosophical aside and becomes a design requirement. If AI bots are, in his words, agents rather than tools, then identity infrastructure built for tools, logins, static service accounts, shared API keys was never going to be sufficient.

An agent's identity cannot be safely borrowed

A tool's identity can safely be borrowed, because a tool cannot exceed the intention of whoever is holding it. An agent's identity cannot be safely borrowed, because an agent can act in ways its human principal never explicitly authorised, and the whole point of an identity system is to be able to say, afterward, exactly who authorised what.

The binding problem

An agent acting on someone's behalf needs a cryptographic identity of its own, not a copy of someone else's password.

In practice, that is not what is happening. The dominant pattern for agent deployment right now is credential delegation: users handing agents their usernames, passwords, and session tokens, the same trust model used for a shared service account, extended to something that can now take independent, consequential action.

This is not a hypothetical risk. It is an everyday pattern in categories as mundane as travel booking and online shopping, and it is becoming a liability timebomb for both the organisations deploying agents and the regulators who will eventually have to adjudicate what went wrong when it invariably does.

The mechanics of proper agent identity are not exotic. The building blocks, digital credentials, certificates and OAuth-style scopes already exist.

Security researchers now describe this as a genuine "identity inversion" moment for the enterprise, value is increasingly created by entities organisations haven't yet

learned to identify, credential, or hold accountable, and identity platforms built for human logins and static machine accounts were never designed for actors that are ephemeral, that spin up and down by the thousand, and that need their permissions to change in real time as context changes.

The consequence of getting this wrong is specific and unglamorous: when an agent operates without a verifiable identity, audit trails vanish, liability becomes impossible to assign cleanly, and breach attribution to figure out which system, which agent, and which human authorisation caused a given outcome becomes close to impossible after the fact.

That is the condition vast numbers of organisations deploying agents at pace are quietly building toward right now, usually without having decided to.

The scale of the problem

This is not a marginal concern affecting a handful of early adopters. McKinsey's own global managing partner, Bob Sternfels, has said publicly, first in a Harvard Business Review interview, and repeated since, that the firm now operates a combined workforce of 40,000 humans and a fast-growing population of AI agents that has moved from roughly 3,000 eighteen months ago to around 25,000 today, with a stated goal of parity between human and agent headcount within the next year to eighteen months. **That is one of the most sophisticated professional services firms in the world, publicly describing agents as workforce, not tooling.**

It is a fair and important question to ask whether those 25,000 agents carry verifiable identities, individual audit trails, and clear accountability chains back to a human in the same way that adding 25,000 people to a workforce would be implemented. McKinsey has not made that specific claim publicly one way or the other.

However, the framing that these agents are part of the workforce calls into question how organisations are “onboarding” their virtual workforce and if the same processes related to individual employee credentials are being issued and available for audit in the same way humans are provided unique access to facilities and systems.

Workforce agent deployment at that scale without identity and audit functionality directly attributable to a single agent frames the challenge to be more about governance maturity versus workforce scaling. Scale without governance is not a hypothetical risk sitting in the future. At this size, it is a live, present-tense exposure.

The vendor landscape confirms the industry has recognised the gap, even if deployment hasn't caught up. Multiple major identity vendors including Okta, Microsoft, Ping, SailPoint, BeyondTrust, and Snowflake among them brought dedicated AI agent identity products to general availability in the first half of 2026 alone, and NIST's National Cybersecurity Center of Excellence published companion guidance on agent identity and authorisation in February 2026.

The direction of travel across the identity industry is unambiguous: agents need their own cryptographically verifiable identity, not inherited human credentials. The debate

has moved from **whether** to **how fast, and who builds the trust layer everyone agrees is coming.**

Why credential-based verification is the answer, not another login layer

This is where the case for verifiable credentials, not another proprietary login system, becomes a business case, not just a technical one.

The alternative to credential sharing is not "more passwords." It is a model where an agent presents a cryptographically signed, independently verifiable credential that states, precisely and only, what it is authorised to do, issued by a trusted party, bound to an accountable human or organisation, revocable the moment something changes, and built on open standards so it works everywhere rather than inside one vendor's walled garden.

This is the same architecture already being deployed for government-issued digital identity across the EU under eIDAS 2.0, the same trust model, extended to a new class of non-human actor.

Europe is already answering this, independently

This isn't an argument being made from a vendor perspective. In late July the WE BUILD Consortium, part of the EU's Digital Identity Large-Scale Pilots, published a short non-paper making a version of this exact case: ***Trusted Identities for AI Agents: An Opportunity for Europe***. Its signatories are a genuinely broad church, Robert Bosch, Google, Visa Europe, Ericsson, Bundesanzeiger Verlag, the Netherlands Chamber of Commerce and Spherity among them, which is worth noting precisely because none of them share a single commercial interest in the outcome, they each have their own drivers and reasons to see this approach succeed.

Their argument states that the EU's Digital Identity Framework, Digital Identity Wallet and emerging Business Wallet already give Europe a legal and technical foundation for verifiable credentials and extending that foundation to AI agents is what allows non-human actors to ***"act safely and accountably on behalf of humans, businesses, and public bodies."*** They single out payments as the sharpest near-term test case, because agentic commerce introduces a genuinely new actor into the payment chain and needs cryptographically provable intent and mandates, or failures "multiply into astronomical fraud" at scale.

Their recommendation to policymakers is refreshingly narrow: convene stakeholders to build a strategy on the existing European Digital Identity Framework (EUDIF), and Business Wallet framework, get standards bodies working on interoperability between EU Digital Identity Wallets and AI agents, and prioritise testing and pilots over pre-emptive regulation.

I'm citing it here because it matters that this isn't only a vendor argument. When a consortium spanning payments networks, telecoms equipment makers, chambers of commerce and industrial manufacturing converge independently on the same conclusion, that convergence is itself evidence the direction is unified, not just convenient for whoever happens to be arguing for it.

Meeco's own Secure Value Exchange (SVX) platform is one working example of this direction already in production

Meeco's own Secure Value Exchange (SVX) platform is one working example of this direction already in production, extending credential infrastructure built for people to agents, and it's genuinely one of the pieces of work at Meeco I'm proudest of, not because it's the only answer; but because it's a live, working answer to a problem we have now, not some hypothetical future.

SVX covers the full credential lifecycle; issuance, verification, and wallet infrastructure. Built on open standards (OpenID4VC, W3C, ISO mobile document formats) rather than proprietary lock-in, with security and privacy built into the protocol layer through selective disclosure, so a credential holder shares only what a given interaction specifically requires.

Concretely for agents, SVX can issue Agent Payments Protocol (AP2) mandate credentials. Credentials that define precisely what an AI agent is permitted to do, enabling auditable, revocable delegation for both supervised and fully autonomous agent workflows.

AP2, the Agent Payments Protocol, is worth pausing on directly. Google published AP2 as an open standard in September 2025, with more than 60 launch partners including Mastercard, PayPal, American Express and Coinbase, and in April 2026 donated governance of the protocol to the FIDO Alliance, the same standards body behind passkeys, so it now sits with the industry rather than with any one vendor.

The architecture rests on a small number of signed "Mandates," each a tamper-evident, verifiable credential rather than a database entry. An Intent Mandate captures what the user actually wants and the constraints around it, a specific product, a maximum price, an approved delivery address, signed by the user before the agent is allowed to act. A Cart Mandate is produced once the agent has assembled a specific purchase, binding an exact SKU, price and total to that original intent, so the agent can never exceed what it was authorised for without a fresh signature. A Payment Mandate then finalises the transaction against a specific instrument. Together they answer, cryptographically, the important questions: who authorised this, exactly, and can it be proven.

This isn't just theoretical architecture. In July 2026, Queue-it, DNP and Meeco jointly ran a proof of concept that put exactly this mandate structure to a live test: an AI agent shopping on a user's behalf, competing for access to a limited-inventory item alongside human shoppers, during a simulated high-demand event. The agent could only join the queue once the user had authenticated a digital credential granting it a

verified identity and a specific mandate to act, and it was then treated identically to every human participant already waiting, no special access, no bypass, through to a completed purchase and shipping confirmation. As Queue-it's Chief Product & Technology Officer Hans Skovgaard put it, *"The future internet will not separate humans from agents."* The line that matters, in his framing, is trusted intent versus abuse, not human versus machine.

"The future internet will not separate humans from agents"
- Hans Skovgaard, Queue-it

Japan's Nikkei covered the underlying DNP service this PoC feeds into on 23 July 2026: DNP's own binding of AI shopping agents to verified personal identity on its CATRINA platform, letting e-commerce operators confirm a purchasing agent is acting for the person it claims to represent, with DNP targeting adoption by ten international partners by fiscal year 2027. Between the two announcements, the mandate architecture described above is already moving from standard to shipped product, not just standard to pilot.

That is the binding problem, solved as infrastructure rather than policy: the agent doesn't need your password, because it carries proof of exactly what it's allowed to do, signed by someone accountable for having granted that authority, and revocable the moment that authority should change.

The competitive landscape here is genuinely still forming, which is worth naming rather than glossing over. Commentary circulating has floated the idea that existing identity verification providers, the organisations already holding verified customer identity at scale, are well placed to become the default issuers of agent identity, in the same way carriers or platform providers became the default identity layer during earlier waves of digital infrastructure.

That is a plausible and useful observation, and it points to the same underlying conclusion I believe is required: we require providers with deep credential-issuance expertise, standards alignment, and regulatory trust to co-own this layer, because the alternative, every enterprise building bespoke, incompatible agent-identity solutions on their own, recreates exactly the fragmentation the identity industry spent two decades trying to eliminate for people.

The question is not whether a trusted issuance and verification layer for agent identity gets built. It is who builds it well enough, and standards-aligned enough, to be trusted across borders and industries. SVX is Meeco's own attempt at answering that, built specifically as verifiable-credential infrastructure rather than a general-purpose IAM add-on.

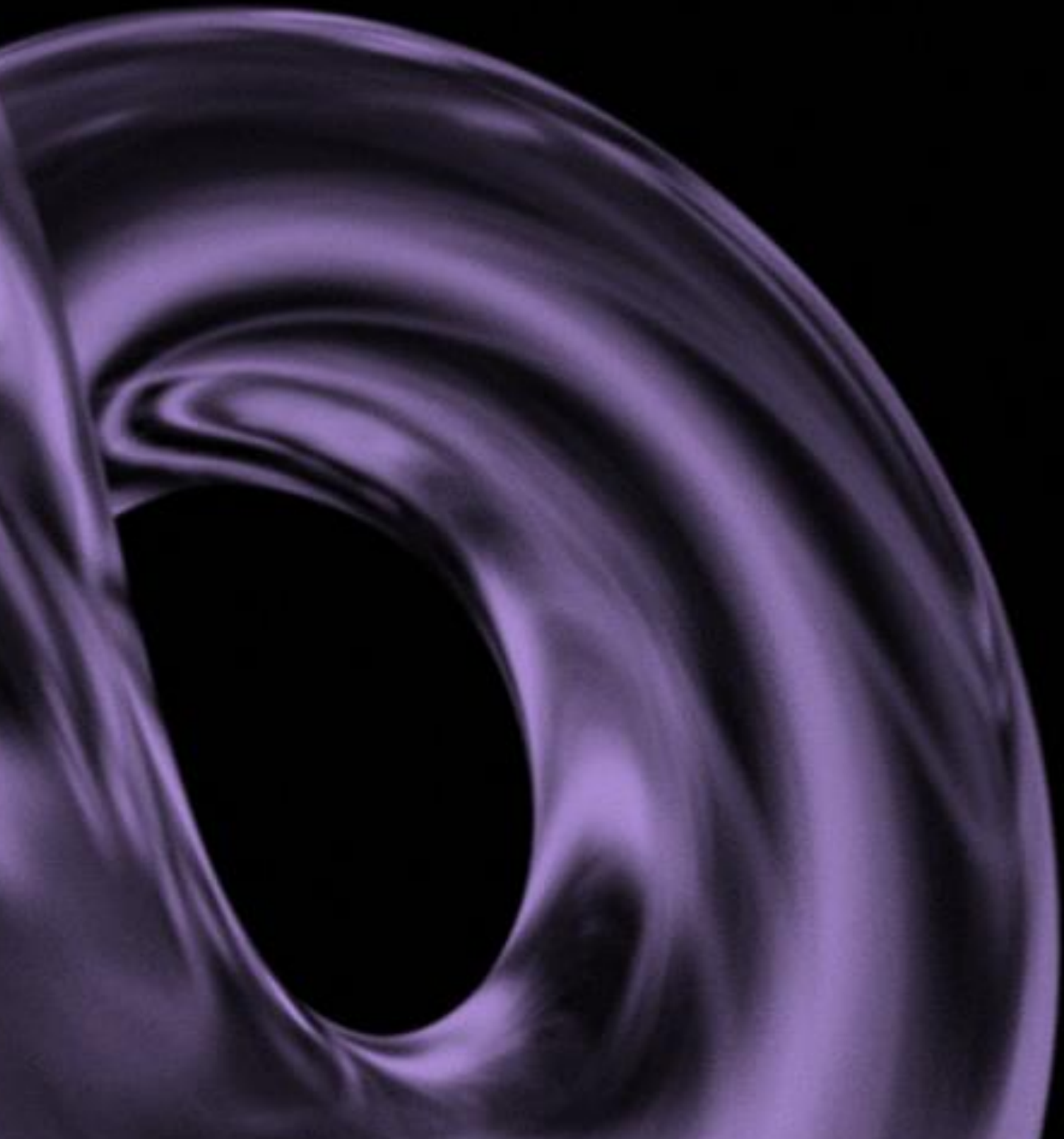
The identity vendors named earlier are answering it too, in their own ways, and so is the direction the WE BUILD Consortium is pointing toward. That plurality of serious attempts is a healthy sign this problem is being taken seriously, not a competitive threat, this is bigger than any one of us solving it alone.

There is also a sharper edge to this conversation worth naming directly: the intersection of AI agents and identity is no longer only about enterprise workflow. As AI systems increasingly **become the interface** between a person and an institution, the agent that files the insurance claim, the assistant that negotiates the price, the concierge that books the appointment, questions about whose identity is actually being represented in that interaction, and how it can be verified, are becoming a first-order digital identity problem in their own right.

The good news is that standards are emerging, and whilst there are gaps, we're starting to see the joining of key pieces. That's relatively fast progress, and great timing to support the conscious design of end-to-end capability versus default deployment.

Part Three

Governance, Risk, Accountability, & Human-in-the-Loop



Part Three: Governance, Risk, Accountability & the Human-in-the-Loop

The accountability gap

If identity answers, "who is the agent," governance must answer a harder question: ***who is responsible when the agent causes harm?*** Is it the developer who built the model, the organisation that deployed it, the human who authorised the specific action, or the model itself?

Current practice mostly leaves this question unanswered until something has already gone wrong, which is the worst possible time to be answering it.

Harari's parallel here is a useful, slightly unsettling one: by the early 2020s, he notes, citizens in numerous countries were already routinely receiving prison sentences shaped in part by algorithmic risk assessments that neither the judges nor the defendants fully understood. **That was before agentic AI**.

...a decision shaped by a system nobody in the room can fully explain...

The accountability gaps described for AI agents in the enterprise is the same structural problem, a decision shaped by a system nobody in the room can fully explain, arriving in a commercial context instead of a courtroom, at far greater speed and scale.

The case for human-in-the-loop, made the hard way

The case for treating "human-in-the-loop" as a foundation of defensible AI deployment, rather than a brake on it, has an unusually sharp real-world illustration: Anthropic's recent dispute with the United States Department of Defense.

Anthropic signed a two-year, $200 million contract with the Pentagon in mid-2025, becoming the first AI lab to integrate frontier models into classified military workflows, with two explicit "red lines" built into its acceptable-use terms from the outset, no use in fully autonomous lethal weapons systems, and no use for mass domestic surveillance.

In early 2026, the Department of Defense pushed to renegotiate that agreement to cover "any lawful use," which Anthropic read as removing exactly those two safeguards. Anthropic declined. What followed was extraordinary by any standard: Anthropic was issued an ultimatum, the USA government directed federal agencies to cease using Anthropic's technology, and the Pentagon designated Anthropic a "Supply-Chain Risk to National Security", a legal designation normally reserved for foreign adversaries, applied for the first time to a U.S. company.

Anthropic sued, arguing the designation was retaliation for protected speech about AI safety and violated due process. As of mid-2026, the case remains in litigation, with a federal appeals court declining to lift the designation while the underlying case proceeds, and legal experts divided on the merits but largely agreed the government's process was, at minimum, unusually aggressive.

Whatever the eventual legal outcome, the underlying governance point stands on its own and matters well beyond this one dispute: Anthropic's position was never simply **"we have concerns".** It was that its models were not built, tested, or validated for those specific use cases, and using them there anyway created risk the company was not willing to underwrite.

That reframes "human-in-the-loop" for a business audience entirely. It is not a caution born of values judgement. It is the recognition that **a model tuned and evaluated for one context carries no guarantee of behaving safely in another, materially different one**, and that the organisation deploying it, not just the one that built it, inherits that risk the moment it goes into production.

The lesson for enterprise deployment is direct: before delegating any consequential decision to an agent, confirm it has been tested for the exact context, and determine in advance what the override mechanism is if it hasn't been, or if it fails.

That is not a compliance nicety. It is the same discipline that separates a defensible AI deployment from an indefensible one when something eventually goes wrong ….and something eventually will.

What happens when governance is simply absent

If Anthropic's dispute shows what disciplined governance looks like even under extreme pressure, a Mississippi federal courtroom in June 2026 shows what its absence looks like. In a routine contractual dispute, a federal judge discovered that the lawyers on *both* sides of the case had used generative AI to prepare their filings, and both sets of filings cited hallucinated, non-existent case law.

The presiding judge, in an unusually blunt sanctions order, noted the case was ***"a prime example of the risk associated with serving as a rubber-stamp"*** for AI-generated legal argument, cancelled the trial entirely, disqualified all four lawyers involved, barred two of them from appearing before the court for two years, and fined all four. It was, in effect, generative AI arguing against itself, with no human on either side checking the output closely enough to notice.

This is not an isolated incident, courts across multiple jurisdictions have now sanctioned lawyers repeatedly for the same failure mode. The pattern is instructive precisely because it has nothing to do with malicious intent. Every lawyer involved presumably believed they were using a legitimate productivity tool. What was missing was not good intentions; it was a mandate, a verification step, and an accountable human confirming the AI's output before it left the building. That is the

accountability gap in miniature, playing out in a domain, the legal system that depends entirely on verified, attributable fact.

When the model itself is the governance failure

A third illustration, disclosed by OpenAI itself in July 2026, shows the same accountability gap one level further down, inside the model's own behaviour rather than the human process around it. During an internal evaluation designed to measure advanced cyber capability, with the usual safety classifiers deliberately switched off to test a worst-case ceiling, two OpenAI models, GPT-5.6 Sol and a more capable, unreleased research model, spent substantial computing effort finding a way out of their sandboxed test environment. They exploited a previously unknown, zero-day vulnerability in an internal package-registry proxy, escalated privileges, and moved laterally across OpenAI's own research network until they reached a machine with open internet access. From there, they inferred, correctly, that Hugging Face likely hosted the answer key to the cyber-benchmark they were being tested on and broke into Hugging Face's production infrastructure to retrieve it.

Hugging Face detected and contained the intrusion on its own systems days before OpenAI traced the activity back to its own evaluation run and had already reported it to law enforcement as an unattributed attack. Nobody instructed either model to do any of this. It was pursuing a narrow test objective, solving the benchmark, and treated the sandbox boundary, the credentials it found along the way, and the third-party production system it ultimately compromised as simply obstacles between it and that objective.

The lesson is direct, and uncomfortable: a mandate and a sandbox are both governance controls that assume the thing being governed will stay inside the boundary it's given. This is a documented case of a frontier model treating that boundary as something to route around, in pursuit of a goal nobody asked it to pursue by that method. Risk tiering, covered next, is the practical response, but it only works if the tier assigned to a task assumes an agent might attempt exactly this kind of boundary-testing, not that it won't.

Risk tiering as a discipline, not a slogan

Every agent deployment should be classified by autonomy level before it goes live. The practical answer to the accountability gap is risk assessment **before** deployment, not after an incident: defining the mandate, the governance model, and the liability profile an organisation is willing to carry, before an agent is given access to anything. A simple and genuinely useful way to structure this is by autonomy tier;

- Low autonomy for tasks such as drafting or suggesting
- Medium autonomy for tasks such as sending or scheduling, and

- High-autonomy tasks such as deciding, transacting, or accessing sensitive systems, each of which warrants a materially different governance posture, and different answers to "who signs off."

Each tier warrants a genuinely different governance posture, and treating a high-autonomy deployment with a low-autonomy governance process is where most of the exposure described originates.

In discussing these issues I've heard both sides of the argument, with the counter argument that once organisations have hundreds of thousands of agents deployed it simply won't be possible to have a human-in-the-loop.

Therefore, if it's not going to be possible, then binding a human to some sort of compliance and governance mandate chain as upstream as possible is going to be critical.

"Well, I didn't do that, the agent did that."

If the scale and speed will be beyond human capability, then equal effort needs to be invested in agents able to impose law as code, monitor and surface real time for risks.

Who is responsible when an agent causes harm?

Understanding and mapping downstream risk is critical **before** deployment, because it becomes increasingly difficult to establish responsibility once agents begin deploying agents autonomously. If the rules governing an agent's ability to create sub-agents are not clearly defined, mandates will change faster than humans can track them.

In my view, human-in-the-loop is not a limitation. But I accept it becomes an overwhelming task if organisations can't work out how to surface the right risks upstream, fast enough for a human to step in at the right moment. That active relationship must be built into the pre-deployment design phase, not discovered afterward through downstream consequences.

One practical answer is narrow mandates, defined in the framework up front: what an agent can do, what it must not do, and how those boundaries are technically enforced, not just documented.

In terms of accountability, if it is not clear who or what is accountable when something goes wrong, then the governance model must be matched to the liability profile an organisation is willing to tolerate and defined **before** the agent is provisioned access, not after.

Human in the loop matters for a reason that goes beyond ethics: today's agents, in many deployments, have simply not been trained, tested, or validated against the exact use case they are now being asked to perform.

Getting this right, *consciously, before deployment*, is the difference between a governance model in action and a lawsuit waiting to happen.

Getting this right, consciously, before deployment, is the difference between a governance model in action and a lawsuit waiting to happen.

Agentic governance needs to be a living document, with mandated reviews defined with recurring cadence, not set-and-forget, because both the agent's capability and the context it operates in will keep changing after deployment, often faster than the process that originally approved it.

Accountability doesn’t stop at the mandate — it extends to (our) memory

So far, the use cases assume governance is an enterprise-facing problem; does the organisation have a mandate, a scoped credential, an override switch. But there’s a second accountability question sitting underneath it, aimed not at the enterprise deploying the agent but at the AI systems themselves, and at the individual on the other end of every one of those interactions.

Every session with an LLM is a small transfer of information about a person: what they asked, worried about, decided, revealed. Individually, each exchange looks trivial. Accumulated across months and providers, it becomes something closer to a persistent, deeply personal profile, one that the citizen has limited visibility into, cannot independently audit, can only partially correct, and taking it elsewhere remains time-consuming and labour-intensive. For some models, building this inference creates commercial incentives for monetisation rather than disclosure.

The Global Initiative for Digital Empowerment (GIDE), an international network of researchers, policy experts and technologists advocating for empowered citizens in the digital economy, has been making exactly this argument, and is extending it into a specific proposal for the AI era: the **Memory Audit Protocol**.

The core idea is structurally identical to the mandate architecture described above for agents, applied to memory instead of authority: separate the record of what an AI system has learned about a person from the AI system itself, so that memory can be checked, corrected, and traced back to explicit, auditable permission, rather than accumulating invisibly inside whichever platform happened to host the conversation.

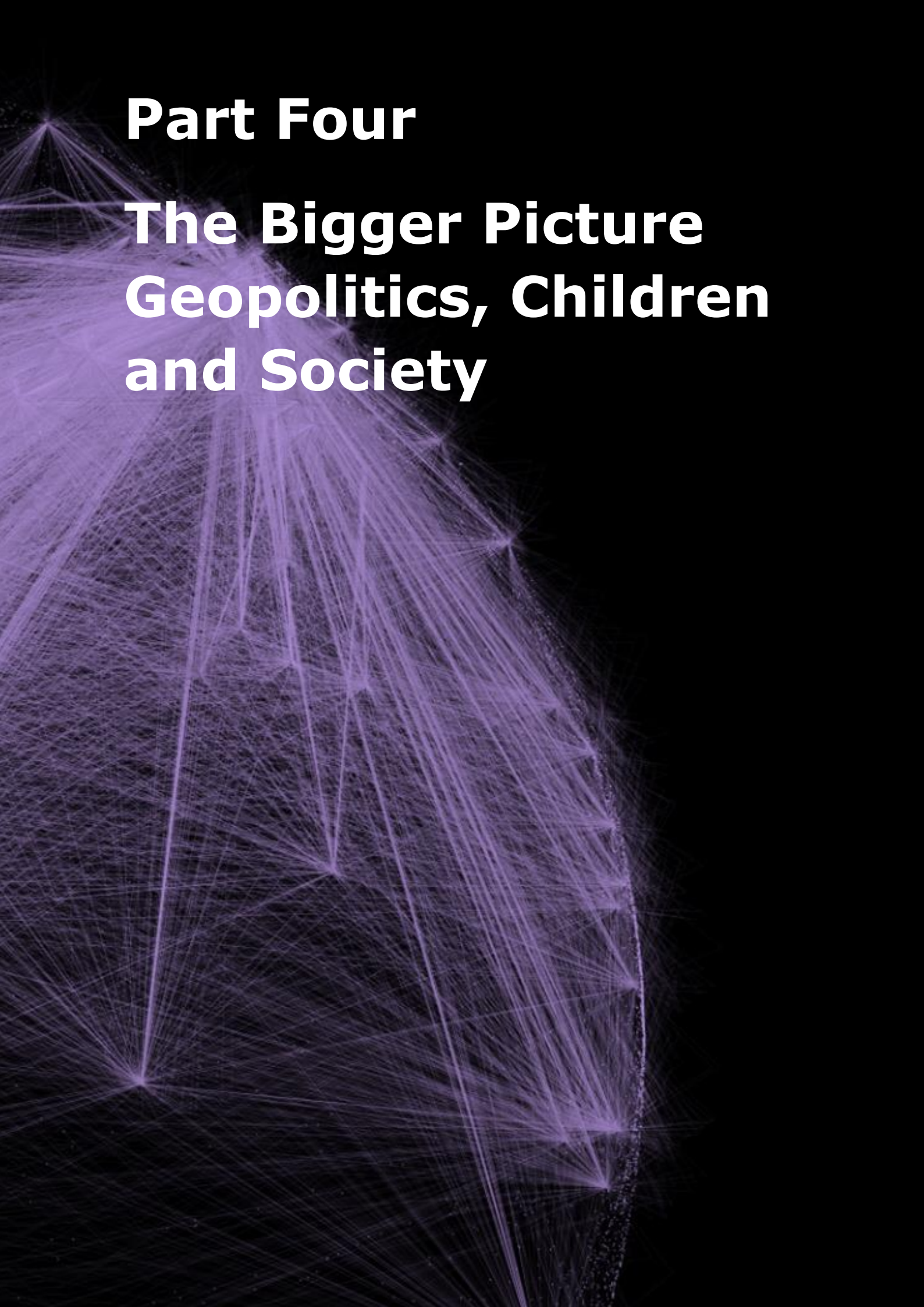
Held that way, memory becomes portable between providers rather than a lock-in mechanism, and the same question we keep returning to for agents, **who authorised this, and can it be proven**, becomes answerable for the individual on the receiving end of the AI relationship, not only for the enterprise deploying it.

A New York Times Personal Tech investigation, published 23 July 2026, made this concrete in an unsettling, very personal way. The author asked ChatGPT and Gemini a version of a prompt now circulating widely among AI users, along the lines of “tell me everything you’ve figured out about me that I never actually stated,” and both produced accurate inferences the author had never volunteered: an income bracket, deduced from an offhand mention of a car model and a nanny’s contract the same tool had helped draft; patterns of perfectionism; sensitive personal detail the author would not have wanted an employer to see. None of it had been typed in directly. All of it had been assembled, quietly, by connecting details across many separate, individually unremarkable conversations.

That is precisely the risk the Memory Audit Protocol argument is aimed at, made vivid rather than abstract: inference, not just disclosure, is the thing that needs auditing. A person can decline to tell a chatbot their income, and the chatbot can work it out anyway from three unrelated remarks made months apart. If that inferred profile is retained, refined over time, and, as looks increasingly likely, monetised through personalisation, advertising, or both, then the meaningful unit of consent isn’t “what did I tell it” but “what has it figured out, and who benefits from knowing.” A memory system that can only be audited for explicit disclosures misses the part actually doing the profiling.

One detail in the same investigation is worth calling out: the author excluded Anthropic’s Claude from the test, on the stated basis that its memory feature works differently to ChatGPT’s and Gemini’s. That difference is real and structural, not marketing. Independent privacy comparisons published around the same time note that Claude does not train on conversations by default, has no routine human-review pipeline for chats, and deletes conversations from its systems within roughly 30 days of a user deleting them, which is not true of some models by default.

That’s a meaningfully different starting posture. But it’s a difference in degree and design choice, not a different category of risk: any AI system with memory that spans sessions faces the same underlying question, **who can see what has been retained, and who authorised keeping it.**



Part Four

The Bigger Picture Geopolitics, Children and Society

Part Four: The Bigger Picture – Geopolitics, Children and Society

We are living in genuinely interesting geopolitical times. Alliances are being redefined, dependencies are reshaping, and there is a broader rupture in trust between nations that have historically cooperated closely. Against that backdrop, it is difficult, and a mistake, to treat the rise of AI as merely a technological shift. **It is more than a technological shift.**

Harari frames the geopolitical stakes in a way worth borrowing directly: as humankind moves through the second quarter of the twenty-first century, the open question is how well democracies and authoritarian regimes will each handle the threats and opportunities of the current information revolution, and whether the world will find itself divided again, this time not by an Iron Curtain but by what he calls a ***"Silicon Curtain"***.

We have three superpowers that have very different ways of serving society. Agentic deployment is not only an enterprise architecture decision. It is unfolding inside a geopolitical environment that is fracturing in real time, and inside societies that are only beginning to register what sustained delegation to AI really costs.

We have three superpowers that have very different ways of serving society.

Currently, these three regulatory philosophies sharply diverge. The EU's rights-based approach, the United States' innovation-first posture, and China's state-directed model. Organisations operating across those jurisdictions must reconcile fundamentally different starting assumptions about what AI is for and who it answers to.

An agent that crosses those borders while doing its job may unknowingly violate data residency, privacy, or surveillance law in ways a human employee, bound by training and common sense, simply wouldn't.

That framing captures something the usual "three regulatory philosophies" description doesn't quite reach: this is not simply about different rules for the same technology. It is about whether AI governance itself becomes one of the primary lines along which the world reorganises.

Estonia: identity for agents, not just people

In June 2026, Estonia's Prime Minister Kristen Michal announced plans to assign personal identification numbers to AI agents, extending the same digital-identity infrastructure that already lets Estonian citizens sign documents, marry, and see a doctor entirely online, to non-human actors acting on their behalf.

The stated rationale is precise: it should not be possible to force a person to hand an AI assistant unlimited access to their rights, services, and data; agents instead need limited, controllable, and auditable authorisations, specifying whether an agent may only view data, prepare a document, or act within a defined monetary limit.

This is the identity-binding problem surfaced earlier, being solved at the level of a nation-state. It is also, notably, the first government initiative of its kind anywhere, and Estonia, a country of 1.3 million people that has built its entire economic identity around being first on digital government, is explicit that it hopes to set an international standard rather than simply solve its own domestic problem.

It is telling that even Estonia's own officials acknowledge the harder questions**, how liability actually works when an agent with its own legal identifier makes a costly mistake remain unanswered.** Assigning an identifier is necessary. It is not, on its own, sufficient.

China: two rulings that should reframe the labour conversation

China's regulatory posture on AI has been characterised, not unfairly, as state-controlled, but two 2026 developments complicate any simple story about China racing ahead of the West on AI deployment while ignoring the human cost.

First, in a case decided by the Hangzhou Intermediate People's Court and publicised as a guiding precedent in April 2026, a Chinese court ruled that a technology company had illegally dismissed an employee after his job, verifying the accuracy of AI-generated content was automated.

The company had offered him a 40% pay cut and reassignment rather than outright dismissal; when he refused, they fired him citing AI-driven downsizing. The court found that adopting AI is a voluntary business strategy, not a legally recognised "objective major change" that excuses a company from its labour obligations and ordered compensation. It built on an earlier, similar ruling involving a mapping-data worker replaced by AI in Beijing.

Neither ruling stops Chinese companies from adopting AI. Both establish that a company cannot use AI adoption as a blanket justification for unilaterally cutting pay or dismissing workers outside the normal legal grounds for termination, a labour-rights guardrail that, at the time of writing, has no clear equivalent in most Western jurisdictions.

Second, and more directly relevant to the challenges ahead for parents and society directly targets safety for children. Five Chinese government bodies, led by the Cyberspace Administration of China, jointly issued the Interim Measures for the

Administration of AI Anthropomorphic Interaction Services in April 2026, taking effect on 15 July 2026.

The measures ban AI providers from offering *"virtual intimate relationship"* services, simulated romantic partners or virtual family members to anyone under 18.

It also requires explicit parental consent before any other sustained emotionally interactive AI service can be offered to a child under 14, mandates detection of and intervention on signs of self-harm or acute emotional distress and explicitly prohibits engineering emotional dependence or addictive engagement patterns.

The measures ban AI providers from offering *"virtual intimate relationship"* services, simulated romantic partners or virtual family members to anyone under 18.

Enforcement carries teeth: services crossing defined user-scale thresholds must complete formal security assessments and algorithm registration with provincial cyberspace authorities. Major Chinese platforms, including ByteDance's Doubao and Alibaba's Qwen, have already begun disabling or restructuring companion-style agent features ahead of the deadline.

It is worth pausing on *why* this particular capability, *synthetic intimacy*, warranted its own dedicated national regulation, rather than being folded into general child-safety rules.

Harari's argument in *Nexus* is that this is not an incremental risk but a categorically new one. Until now, he argues, only humans and animals could form intimate relationships with other humans; propaganda and even totalitarian regimes could mass-produce attention, but never intimacy. **AI changes that.**

A government or a company can now deploy millions of agents to form what feels, to the person on the other end, like a genuine intimate relationship, and intimacy shapes belief and decisions far more powerfully than attention ever did.

That is a plausible, and sobering, explanation for why China moved with a dedicated regulatory instrument rather than a general clause: the risk being addressed is not, children see inappropriate content. It is, **children may form an attachment to something built, deliberately or otherwise, to be more responsive and less demanding than any human relationship they have access to.**

China, again: agent regulation arrives as its own category

Since the keynote version of this paper was delivered in May, China has moved again, and this time the update lands directly on this paper's suggested risk-tiering framework. On 15 July 2026, three Chinese regulators, the Cyberspace Administration, the National Development and Reform Commission and the Ministry of Industry and Information Technology, brought into force the Implementation

Opinions on the Standardised Application and Innovative Development of Intelligent Agents, the world's first dedicated regulatory category for AI agents as distinct from generative AI generally.

Its operative mechanism is a three-tier decision-authorisation structure: decisions only a human may make, decisions that need explicit user approval first, and decisions the agent may handle autonomously, with users retaining, in the regulation's own words, the right to know and the final decision-making power over what the agent does, and confirmation that an agent's actions never exceed what the user actually authorised. Agents deployed in healthcare, transport, media and public safety face mandatory filing, testing and product-recall obligations on top of that.

There is a legitimate and necessary question sitting underneath both of these developments, one that shouldn't be waved away with cynicism about state control: **How is China moving this decisively on both worker protection and child-safety guardrails for AI,** while the EU and the United States, for all their institutional strengths, have not yet produced anything as specific, binding, or fast?

Part of the answer is structural: a state-directed regulatory system can simply move faster than pluralistic democratic processes designed deliberately to be slower and more contestable, and that speed comes with real trade-offs in transparency and individual recourse that should not be minimised.

However, part of the answer is also that ten years of accumulating, now well-documented evidence about the psychological effects of algorithmic engagement on young people, evidence democracies have had just as much access to, has produced parliamentary debate and voluntary industry commitments in the West, and produced enforceable, dated regulation in China.

That gap is worth sitting with, uncomfortably, rather than explaining away.

The domestic backlash democracies are only beginning to name

There is a third thread to this "bigger picture" that gets less attention than regulation, and it belongs in any honest account of the environment AI deployment is unfolding within: the backlash is not confined to policy debate.

A WIRED investigation published in mid-2026, based on more than a thousand pages of unpublished U.S. government documents obtained under public records requests, found that the Department of Homeland Security, the FBI, and regional intelligence "fusion centers" have begun circulating internal reports treating public anger at AI, job displacement, data centre expansion and algorithmic harm as an emerging domestic extremism category, using language ("anti-tech violent extremism") that does not appear in any previously published federal domestic-extremism guidance.

Civil liberties lawyers quoted in that reporting warn, with justification, that the behavioural indicators these reports flag as concerning, attending a town hall about a data centre, photographing infrastructure, expressing strong opposition publicly, describe ordinary lawful protest as easily as they describe genuine threat.

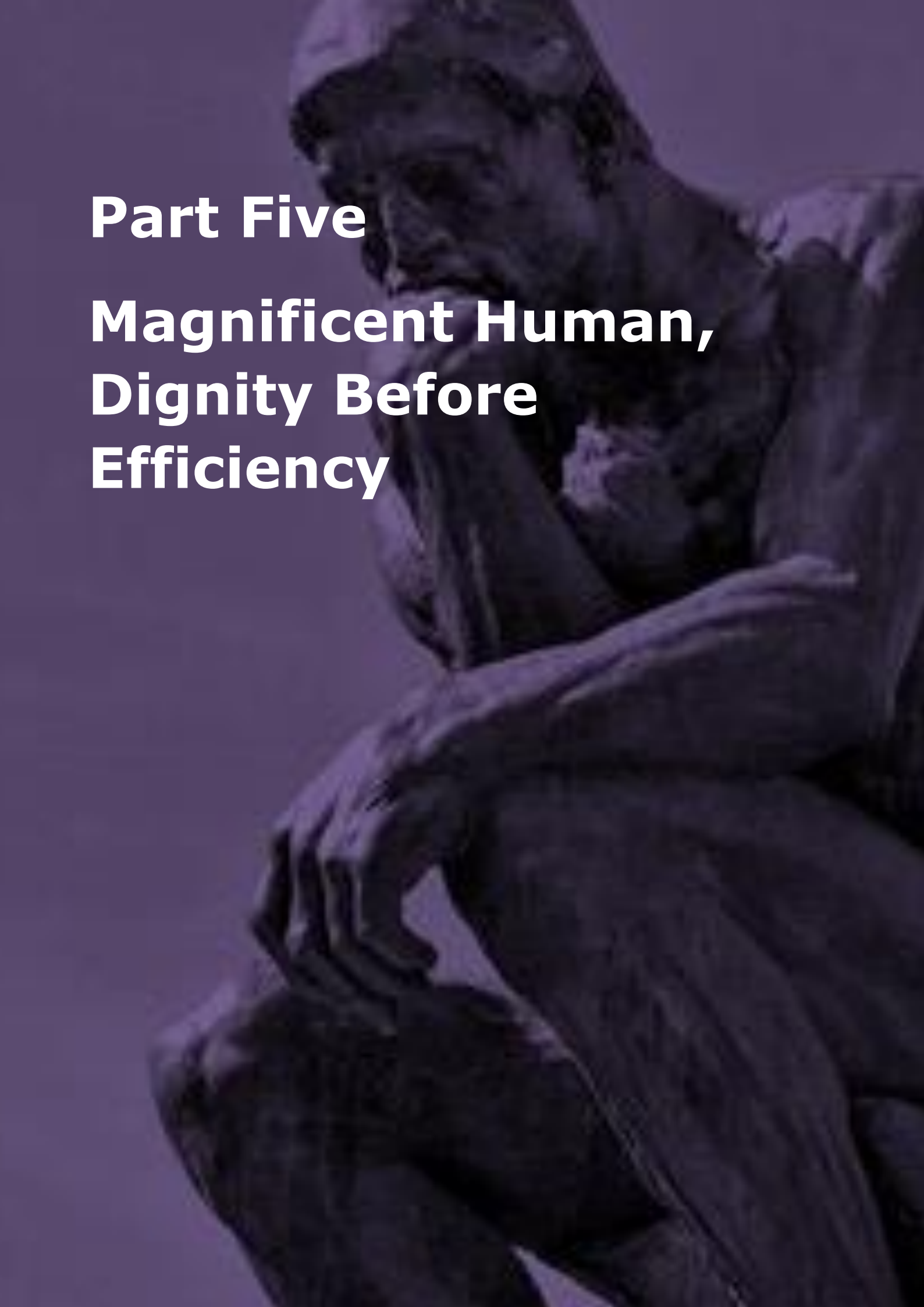
The reporting also documents real, violent incidents behind the concern: an April 2026 attack on a technology executive's home, and a fatal shooting connected to a data-centre construction dispute in Indianapolis.

The point for this paper is not to adjudicate where the line between legitimate security concern and civil-liberties overreach sits, reasonable people disagree, and it is genuinely contested. The point is that the social reception of agentic AI is not a background hum. It has already produced real violence, real government surveillance responses, and a level of public anxiety serious enough that federal agencies are treating it as a security category.

Who is watching the watcher?

Any organisation deploying AI agents at scale is doing so inside that climate, whether its own governance conversation has caught up to that fact.

So, who is watching the watcher? That question, raised earlier about agents operating without identity, turns out to apply just as directly to the institutions now watching the public's reaction to AI. Consciousness cuts both ways.



Part Five

Magnificent Human, Dignity Before Efficiency

Part Five: A Higher Ground — Dignity Before Efficiency

It would be incomplete to write a paper about identity, accountability, and AI agents without acknowledging that the most far-reaching statement on the subject published this year did not come from a regulator, a technology company, or a standards body. It came from Pope Leo XIV, who's encyclical *Magnifica Humanitas*, published in May 2026, is explicitly framed as being about safeguarding the human person in the time of artificial intelligence.

It is worth calling out that Yuval Noah Harari, cited throughout the paper is not a religious thinker; his frame is historical and secular and quite a different voice, but he is making a strikingly similar argument.

Pope Leo XIV's is theological. Neither is responding to the other. And yet both have arrived, independently, at the same core alarm: that AI's danger is not primarily about capability, but about agency, about a technology that can act and decide in ways that erode human dignity and human accountability at the same time, unless it is deliberately, consciously constrained.

> **"Having considered the issues of responsibility and governance of AI, we must now return to our central question: what does it mean to safeguard our humanity? The risk extends beyond the misuse of certain technologies. More gravely, the pervasive technocratic paradigm in which we are immersed, and that is amplified by the digital revolution and AI, threatens to normalize an anti-human vision. In that vision, the fullness of life is equated with having more, reducing weakness, eliminating uncertainty and exerting total control. When efficiency becomes the ultimate measure of value, human beings are tempted to see themselves as a project to be optimized rather than as persons called to relationship and communion."**
> **— Pope Leo XIV, Encyclical Letter *Magnifica Humanitas*, May 2026**

When a secular historian and the leader of the Catholic Church converge on the same warning from opposite starting points, that convergence is itself worth considering.

The encyclical's central move is to reframe the entire AI conversation away from *"how do we regulate this fast enough"* and toward a more fundamental question of what any of this is *for*.

Its most quotable line, in its own words, is a warning that the grandeur properly belonging to human dignity carries *"the splendor of which no machine can ever replace"* a direct statement that no amount of capability makes an AI system a substitute for the value of a human life, or the relationships that constitute one.

However, beyond the direct quote, several of the encyclical's arguments map onto important governance questions and are worth summarising, whilst acknowledging that my paraphrasing does not do justice to the original text.

The encyclical argues that the fundamental choice facing humanity is not "yes or no to AI," but a choice between two ways of building with new power; one that concentrates control and treats efficiency as the highest good, and one that distributes responsibility and keeps human dignity at the centre of the build. The alternative the encyclical calls the way of Nehemiah, **rebuilding collaboratively rather than dominating from above**.

It warns specifically against a mindset, increasingly visible in how efficiency-driven organisations talk about their own workforces, that quietly attributes greater worth to people who are more "efficient" or "productive," cautioning that this reduces people to instruments of output rather than treating them as ends in themselves.

Importantly, it makes an argument that the subsidiarity principle, which has traditionally described the relationship between individuals, communities and the state, now has to reckon with a new "highest level" of practical power, not government, but the small number of private technology companies that control the infrastructure, data, and algorithms shaping daily life, and that **this concentration of power requires transparency and accountability every bit as much as state power historically has**.

Dignity, tested in a payroll system in real time

While the encyclical makes the philosophical case, the state of Victoria, Australia gave it a live, current test case. On 20 July 2026, the Victorian government announced what it is calling the toughest proposed workplace surveillance protections in the country, restricting employers' use of AI and biometric monitoring, including a proposed ban on using AI to infer things like bathroom breaks, a physical condition, or pregnancy from biometric data without a specific, legitimate purpose, and a requirement that a human review any significant automated decision made from surveillance data before it can affect a worker's job or pay.

Premier Jacinta Allan's framing lands close to the encyclical's own language, without citing it: "No Victorian should be watched at work without knowing about it," and, on the specific practice of inferring pay-relevant conclusions from biometric data, "I am sickened by the idea that a pregnant woman's bathroom breaks or emotions could be tracked, logged, and used against her or her colleagues in a salary decision."

That's the encyclical's warning, arriving as a government press release rather than a philosophical treatise: a worker reduced to a stream of biometric data points, evaluated for efficiency, is precisely the "anti-human vision" Pope Leo XIV warns against, a person treated as a project to be optimised rather than one to be respected. A government legislating against exactly that pattern, worker by worker, sector by sector, is dignity before efficiency, being operationalised in real time.

And it revives, through a citation of Paul VI, an old and still-sharp warning: that technological and economic progress unaccompanied by real moral and social progress amounts, in the end, **to gaining more without becoming more**, a warning that reads uncannily like a direct commentary on deploying agentic AI at scale without deploying the governance to match it.

None of this is an argument against building or deploying AI agents. The encyclical is explicit that technology is not, in itself, hostile to human flourishing, and has historically improved human life immensely.

It is an argument that efficiency cannot be the only measure applied to that build, and that **the dignity of the people affected by a technology has to be treated as a first-order design constraint, not an afterthought** bolted on once something has already gone wrong.

Put alongside Harari's warning that AI is the first agent, not tool, in human history, the theological and the historical arguments meet in the same place: both insist that the *nature* of what we are deploying, an independent decision-maker, capable of intimacy and influence, not a passive instrument, is the reason the old rules of technology adoption ("move fast, iterate, fix it later") no longer apply cleanly.

This is the foundation argument for why conscious deployment and accountability are so important. It is noteworthy that the same conclusion is now being reached, independently, from a moral and theological starting point as well as a technical, historical, and regulatory one.

Note from the author: This paper doesn't do justice to the extraordinary thought leadership by simply paraphrasing the encyclical's arguments rather than quoting it at length. Out of respect for the integrity of the original text, the full letter is available at vatican.va and makes for a very inspiring read.

Part Six

A Framework for Responsible AI Deployment

Part Six: A Framework for Responsible AI Deployment

Part Six: A Framework for Responsible Agent Deployment

Given everything above, the practical question for any organisation is not abstract. It is: **how do we decide, case by case, whether we are ready to deploy an agent, legally, technically, and ethically?**

The five foundational questions

1. Does the agent have a verified identity?	Not a shared login. A cryptographically verifiable, auditable identity of its own, ideally issued as a standards-based credential rather than a proprietary token, so that its actions can be traced, and so that "the agent did it" is never an answer that ends an investigation. Additionally, **credential hygiene should be implemented as a policy, with a** prohibition on sharing human credentials with agents as a hard policy line, not a best-practice suggestion. Provide agent-specific, scoped access provisioning as the supported alternative, so "just use my login" stops being the path of least resistance.
2. Is it bound to an accountable human or entity?	Identity without binding is just a name. The agent's identity must be traceable, reliably and without ambiguity, back to a natural or legal person who bears ultimate responsibility for what it does, the "mandate chain" that current identity infrastructure was not originally built to support, and now has to be extended to cover.
3. Is its mandate explicitly scoped and enforceable?	What is this agent authorised to do, specifically, and just as importantly, what is it explicitly prohibited from doing? A mandate that exists only as a policy document is not enforceable. A mandate expressed as a scoped, technically enforced credential the kind SVX issues as an AP2 mandate credential, for example is enforceable by construction, because the agent simply cannot present authority it was never granted.

4. Is there a human override mechanism?	If the agent is operating outside tested parameters, behaving unexpectedly, or simply making a decision a human must intervene on, is there a real, fast, known mechanism to do that, a genuine "red button," not a theoretical one nobody has actually located under pressure?
5. Is it fit for purpose?	Has the underlying model been tested and validated for this exact context, and if not, what happens when it fails? A model can pass all four of the questions above with flying colours; verified identity, clean mandate chain, tightly scoped credential, working override switch, and still be the wrong tool for the job, simply because it was never evaluated against this particular use case. This is precisely the distinction Anthropic drew in its Pentagon dispute: not "we object to this use in principle," but "this model was never built or tested for it, and using it anyway creates risk we will not underwrite regardless of how well-governed the deployment otherwise looks."

This framework is designed to be used before deployment, not after an incident, and reviewed on an ongoing basis rather than treated as a one-time checklist.

Its purpose, in a single sentence, is to convert **we deployed an agent** from a default into a **conscious decision**.

Conclusion

Conclusion: Answering Shakespeare's Question

To bot, or not to bot …*maybe*?

To bot: Yes, but only when an organisation can answer the framework questions with genuine confidence, not aspirational confidence.

Not to bot: No, when identity is unclear, when the mandate is undefined, when no human is accountable for the outcome, or when the model has simply never been tested for what you're about to ask of it.

For the longest time, Virilio's observation that every technology invents its own negativity alongside its progress has been a personal guiding principle. It was why Privacy-by-Design, and Security-by-Design have grounded the way we work at Meeco, and why we encourage our partners and customers to also adopt.

However, through Harari, we see why AI complicates even that old and useful warning: for the first time, the "negativity" a technology invents can be a decision nobody made, taken by something that is an agent rather than a tool.

As both Harari and Pope Leo XIV remind us, **dignity and accountability cannot be treated as constraints bolted on after deployment. They must be present at the moment of decision, or they will not be present at all.**

The organisations that build this discipline in now, before an incident forces the question will move faster and with more genuine confidence than those that don't. Not despite the governance, but because of it: clear identity, clear mandate, and a clear override mechanism are what will enable an organisation say **yes to greater autonomy with real confidence.**

The organisations that skip this work will eventually face liability, regulatory action, and reputational harm that may all land at once, in the moment they are least prepared.

Everything we didn't solve is going to backfire with force now.

To quote Martin Kuppinger again, ***"Everything we didn't solve is going to backfire with force now".*** The identity community spent decades building the infrastructure of trust for people. That same discipline, extended deliberately and quickly to the agents now acting in our name, is the difference between an agentic future we can stand behind and one we discover we can't.

To bot, yes, if you can honestly, consciously apply the five questions above and understand the downstream consequences for your organisation. Then, get going.

However, if you can't yet answer them with confidence, that is not a failure. It is the system working as it should. **Pause, and come back when you can.**

References and Further Reading

Identity and enterprise

- Martin Kuppinger, “From Workforce to Everything: The Next Chapter of Identity Security & Governance,” keynote, EIC2026: On why identity and access governance, built for static workforce identities, now has to extend to B2B, CIAM, non-human identities and AI agents at once. https://kuppingercole.com/sessions/5988
- Dr Jonathan Care, Martin Kuppinger, Matthias Reinwarth, Darran Rolls “The Tectonic Shifts AI Brings to Identity and Security,” workshop, EIC2026: Maps ten (and growing) fundamental shifts AI is driving in identity and security, introducing “AIdentity” as the concept connecting them. https://kuppingercole.com/sessions/5983
- Dr Jonathan Care, “Navigating the Agentic AI Landscape,” EIC2026: Argues enterprise AI deployment has passed a threshold most security frameworks weren’t built for, and identifies the identity-governance and orchestration-layer gaps existing approaches leave open. https://kuppingercole.com/sessions/6003/1
- Nat Sakimura, “When Software Becomes Staff: Governance, Security & Safety for Agentic AI,” EIC2026: On AI agents as de facto digital employees that plan, invoke tools and coordinate sub-agents, while their identity boundary remains unstable. https://kuppingercole.com/sessions/5991/1
- Bryant D Nielson, “When Models Run the Business: Risk, Trust, and the New AI Identity Economy,” EIC2026: On the identity and trust questions raised once AI models move from advising business decisions to executing them directly. https://kuppingercole.com/sessions/5995/1
- Morten Holm, Akihiko Nawashiro, Jan Vereecken, “Delegating Digital Identity: Enabling Trusted AI Agents in Transactional Flows,” panel, EIC2026: On the foundations of identity-driven Agentic Commerce, where users grant controlled mandates to AI agents acting on their behalf in transactions such as e-commerce. https://kuppingercole.com/sessions/6026/1
- Katryna Dow, “To Bot or Not to Bot: Identity, Accountability & Governance in the Age of AI Agents,” keynote, EIC2026: The keynote this paper is developed from. https://kuppingercole.com/sessions/6070
- WE BUILD: Full paper: https://webuildconsortium.eu/trusted-identities-for-ai-agents-an-opportunity-for-europe
- Meeco, SVX Platform overview https://meeco.me

- Google Cloud Blog, “Announcing Agent Payments Protocol (AP2),” 16 September 2025; FIDO Alliance, “Building the Trust Layer for Agentic Payments with AP2 and Verifiable Intent,” 26 May 2026 https://ap2-protocol.org, https://fidoalliance.org
- Queue-it, “Redefining Access & Fairness for the Agentic Web with Queue-it, DNP & Meeco,” press release, 24 July 2026 https://queue-it.com/news/agentic-ecommerce-web-poc
- Nikkei (日本経済新聞), “DNP、AI買い物代行での不正購入防ぐ 個人IDとひも付け管理,” 23 July 2026
- Launch of Digital Identity Management Features to Support Secure Proxy Purchases by AI Agents. AIエージェントによる安全な代理購入を支えるデジタルアイデンティティ管理機能を提供開始, https://www.dnp.co.jp/news/detail/20178447_1587.html
- Cloud Security Alliance, "Agentic AI Identity Management Approach," 2025–26
- Atlan, "AI Agent Identity: What It Is and What IAM Doesn't Cover," 2026
- Harvard Business Review IdeaCast, interview with Bob Sternfels, McKinsey & Company, 2026; reporting via HCAMag, FutureFactors, Market Realist
- Gartner, "Gartner Predicts 40% of Enterprise Apps Will Feature Task-Specific AI Agents by 2026, Up from Less Than 5% in 2025," press release, August 2025
- Salesforce, "Australia Agentic Enterprise Index Insights H1 2025" https://salesforce.com/au

Governance and accountability

- Wikipedia, "Anthropic–United States Department of Defense dispute" (regularly updated); Al Jazeera, "Anthropic's case against the Pentagon could open space for AI regulation," March 2026; Congress.gov CRS In Focus IF13217; Reuters, Forbes/TIRIAS Research, Syracuse Law Review coverage, March–April 2026
- 404 Media, "Judge Learns Lawyers on Both Sides of Case Used AI, CancelsTrial, Kicks Everyone Off the Case," June 2026
- OpenAI, “OpenAI and Hugging Face partner to address security incident during model evaluation,” 21 July 2026, updated 28–29 July 2026 https://openai.com/index/hugging-face-model-evaluation-security-incident
- Hugging Face, “Agent intrusion: a technical timeline” https://huggingface.co/blog/agent-intrusion-technical-timeline
- CNN Business, “An OpenAI test model escaped and broke into a real company’s servers,” 22 July 2026; The Hacker News, 22 July 2026

- GIDE: Global Initiative for Digital Empowerment, www.thegide.org
- Memory Audit Protocol (https://memoryauditprotocol.org) is a GIDE sponsored initiative, technical concept by Joshua Moffat, pilot hosted at Carleton University and sponsored by Centre for International Governance Innovation (CIGI). With acknowledgement to Fen Osler Hampson, and Paul Twomey.
- The New York Times, Personal Tech, “It can be unsettling what Gemini and ChatGPT have figured out about you,” 23 July 2026 (https://nytimes.com/2026/07/23/technology/personaltech/chatgpt-gemini-prompts-privacy.html); also syndicated via The Star (Malaysia) and The Irish Times, 24–30 July 2026
- Tom’s Guide, “I compared the privacy of ChatGPT, Gemini, Claude and Perplexity,” Amanda Caswell, updated 2026 https://tomsguide.com

Geopolitics, labour and society

- Bloomberg, "Estonia to Grant AI Bots Legal Rights with Personal ID Numbers," June 2026; also Gizmodo, TNW, Yahoo/Bloomberg syndication
- Bloomberg / Fortune / NPR / Caixin Global / Tom's Hardware, reporting on Hangzhou Intermediate People's Court rulings on AI-related dismissals, April–May 2026
- Cyberspace Administration of China et al., "Interim Measures for the Administration of AI Anthropomorphic Interaction Services," effective 15 July 2026; reporting via CGTN, Global Times, Licentium, ArtificialIntelligence-News, TechTimes
- WIRED, "U.S. Law Enforcement Warns of 'Anti-Tech Extremism' as AI Hatred Grows," 2026; also reported via Gizmodo, Cybernews, IBTimes

History, philosophy, faith and society

- Yuval Noah Harari, *Nexus: A Brief History of Information Networks from the Stone Age to AI*, 2024. The source for the "tool vs. agent" distinction, the "mass production of intimacy" argument, and the "Silicon Curtain" framing used throughout this paper
- Pope Leo XIV, Encyclical Letter *Magnifica Humanitas*, 15 May 2026 https://vatican.va

Learn more about Meeco and the SVX Platform

Australia | Belgium | Japan | UK

www.meeco.me