Agentic AI, trust and architecture

Thematic paper

JULY 2026

WEB4HUB: 'A SPACE FOR THE METAVERSE – VIRTUAL WORLD AND THE TRANSITION TO WEB 4.0'

European Commission

PPMi
Part of the Verian Group

WEB4HUB

Authors: Rūta Gabaliņa, Elene Seturidze, Oleksandra Yevdokymova, Aida Zepcan.

Acknowledgements: Antoine-Alexandre André, Martina Barbero, Maryna Butym, Egidijus Barcevičius, Pierre Marro, Hubert Romaniec, Fabrizio Sestini, Dimitrios Thomas and others who contributed through interviews, workshops or reviews of the present paper.

Contents

Introduction **4**
Objectives of the paper 5
Problem definition: how agentic AI creates challenges for the existing trust frameworks **6**
Trend A: agentic impact on collective intelligence and digital commons **7**
Effects on trust and architecture 8
Emerging solutions 8
Implications and opportunities to capture 9
Trend B: agentic search and web browsing replace click-based discovery **11**
Effects on trust and architecture 11
Emerging solutions 13
Implications and opportunities to capture 14
Trend C: internet security architecture adapting to agentic AI **15**
Effects on trust and architecture 15
Emerging solutions 17
Implications and opportunities to capture 18
Trend D: on-device agents turn the operating system into a gatekeeper **20**
Effects on trust and architecture 20
Emerging solutions 21
Implications and opportunities to capture 22
Trend E: an agent-native protocol layer forming atop HTTP **24**
Emerging initiatives 24
Effects on trust and architecture 25
Implications and opportunities to capture 26
Trend F: discovery mechanisms emerging for A2A communication **28**
Effects on trust and architecture 28
Emerging solutions 29
Implications and opportunities to capture 31
Trend G: the rising challenge of tracing agents and their actions **33**
Effects on trust and architecture 34
Emerging solutions 35
Implications and opportunities to capture 36
Trend H: provenance layer shaping for agent-generated content **37**
Emerging solutions 37
Effects on trust and architecture 39
Implications and opportunities to capture 39
Trend I: agent-mediated commerce and the future of payment infrastructure **41**
Effects on trust and architecture 41
Emerging solutions 42
Implications and opportunities to capture 43
Key findings **45**
References **48**

List of figures

Figure 1. The agentic paper trilogy 5
Figure 2. The nine trends 5
Figure 3. Machine load vs human engagement 7
Figure 4. Early signs of the shift towards agent-mediated discovery 11
Figure 5. Evolution of web discovery from search to action 12
Figure 6. Security threats specific to agentic AI 15
Figure 7. On-device agents move the trust boundary inward 20
Figure 8. Emerging agent protocol stack 24
Figure 9. Agent identity vs agent tracing and accountability 33
Figure 10. Provenance mechanisms in agentic web 37
Figure 11. Trends' impact and the EU's leverage 45

List of boxes

Box 1. OpenClaw and the limits of cybersecurity in an agentic web 17
Box 2. DNS-AID description 29

List of tables

Table 1. Emerging security mitigations 18

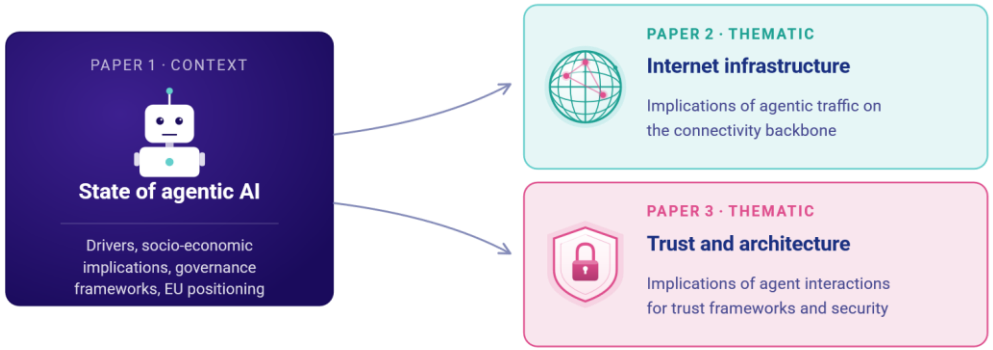
Introduction

The spread of agentic artificial intelligence of agentic artificial intelligence (AI) marks a step change in how the web works. It is shifting everyday online activity from direct human interactions towards more and more tasks carried out by autonomous software agents. This paper examines what that shift means for the future web: how people, organisations and public authorities can retain meaningful control when decisions and transactions are increasingly delegated to agents that act on users' behalf and interact with each other at machine speed. It focuses on the conditions under which agent-mediated environments can remain trustworthy and accountable, looking at **data governance**, **trust frameworks**, **internet architecture** and **security** in an agent-populated web. The paper invited for a further discussion on the kinds of rules, protocols and infrastructures needed to keep the emerging agentic web aligned with public values.

The timing of this shift is critical because the **underlying technologies are moving from experiments to deployment,** and key building blocks of the next phase of the web are already being defined. As agents become **embedded in browsers, operating systems, payment rails and cloud platforms**, early architectural and governance choices are quickly becoming irreversible. This is therefore a decisive moment both for the evolution of internet architecture and for the EU. Decisions taken in the coming years about identity, delegation, tracing, discovery and security for AI agents will shape who can set the rules of an agent-populated web, whose infrastructures others must rely on, and whether the resulting ecosystem reflects European commitments to openness, privacy, competition and fundamental rights. Likewise, they can also lock in new dependencies that are difficult to unwind. By setting out the main trends, risks and emerging solutions, the paper aims to inform timely action on standards, regulation and investment, while there is still a realistic window to influence how agentic AI is built into the next generation of the web.

This paper builds on a comprehensive research and stakeholder consultation exercise carried out under the project **"Web4Hub: A Space for the Metaverse – Virtual Worlds and the Transition to Web 4.0"**, implemented by PPMI and TNO on behalf of the European Commission. The analysis draws on an extensive review of scientific and policy literature, expert interviews, and insights from project activities. The paper is part of a **trilogy on agentic AI** prepared within the WEB4HUB project, examining the technology's implications for the future of the web. The first part, "Agentic AI: State of play", introduces agentic AI in detail, examines its broader societal and economic dimensions, and explores possible futures for Europe's positioning in the agentic era. It also provides important **definitions of AI agents, agentic AI and agentic web**, elaborating on the differences between them. The second part of the series focuses on the physical and operational infrastructure through which agentic systems must function. This paper addresses the implications for data governance, trust frameworks, internet architecture, and security. Together, the three papers cover the principal dimensions of the agentic transition: the socio-economic context and strategic choices; the infrastructure prerequisites; and the governance of data, trust, and security in an agent-populated web.

Figure 1. The agentic paper trilogy



Together: a coordinated examination of how agentic AI is reshaping the future of the web

Objectives of the paper

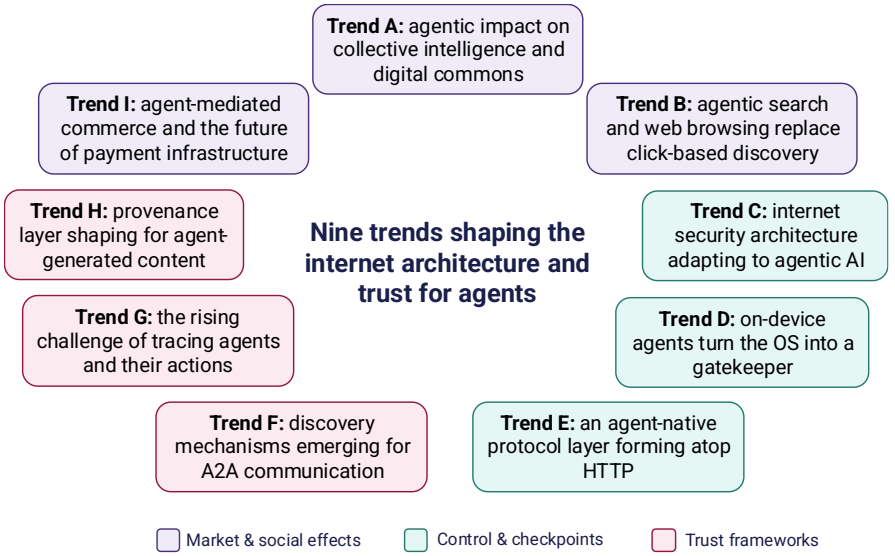
The increasing use of autonomous agents across everyday online services raises fundamental questions about how trust is created, maintained and enforced on the web. This paper sets out to clarify what an agent-populated internet requires from its trust and data governance arrangements, and whether current frameworks and architectural choices are ready to support it. Drawing on research and stakeholder input, the paper aims to provide policymakers, regulators and technical communities with a structured view of the main challenges and emerging design options, so that regulatory initiatives, standards processes and public investments can support trustworthy agentic AI while preserving openness, privacy and fundamental rights.

The paper addresses **three core research questions**:

- What new demands does the rise of agentic AI place on trust, data governance and internet architecture?
- Which solutions are emerging to meet these demands, and how mature and scalable are they in practice?
- What actions should policymakers, standards bodies and infrastructure operators take to ensure that agentic AI develops in line with European values and public interest?

The analysis that follows is organised around **a set of observable trends** in how agentic AI is reshaping trust and architecture on the web. Each trend examines a specific aspect of the agentic transition, combining evidence from technical developments, emerging standards, commercial practice and expert insight. Taken together, these trends can be grouped into three broad categories. The first concerns **trust frameworks**, asking whether agents can be reliably identified, traced and verified. The second focuses on **control and chokepoints**, examining where power and dependency concentrate as agents are embedded into operating systems, agent-native protocols and security architectures. The third addresses **market and societal effects**, looking at who captures value and impact in agent-mediated web through areas such as agent-mediated commerce, agentic search, and the impact on collective intelligence and digital commons. See the mapping of these trends in the figure below.

Figure 2. The nine trends



Problem definition: how agentic AI creates challenges for the existing trust frameworks

Trust on the internet has long depended on a workable link between content, actions and identifiable actors. The existing digital governance frameworks are primarily designed around human users[1]. We may not always know who sits behind a screen, but the web's basic architecture still assumes that messages, transactions and decisions can be tied to a person, an organisation or a clearly accountable system. The spread of agentic AI starts to unsettle that assumption because **software can now plan, decide and act** across digital environments with limited direct supervision.

This fundamentally changes the nature of trust online. Earlier generations of AI operated within an environment where users mainly had to judge whether information was accurate and reliable, supported by built-in web mechanisms signalling whether specific webpages could be trusted. Today, agentic systems introduce a more demanding form of trust: confidence that agents can safely and appropriately act on users' behalf[2]. As agents gain access to communication channels, accounts, and software tools, **failures or manipulation** can lead to unauthorised actions, privacy breaches, security incidents and other real-world harms[3].

The shift from generating content to taking action matters because **trust online is no longer confined to judging whether information is true**. Once agents can send messages, make purchases, schedule services, negotiate terms or trigger other systems, trust also depends on whether those actions can be traced, verified and challenged[4]. The risks compound as agents start operating across several services and steps at machine speed, which makes failures harder to detect and easier to scale. This shift is not limited to the network layer, as agents embedded in devices and connected objects can initiate actions that never pass through conventional online interfaces. Trust therefore moves towards the intermediary layer that governs how agents access data, interpret instructions and execute actions within digital environments and on the edge. This layer amounts to a new gateway to the internet: rather than navigating from link to link, users increasingly interact through an agent that selects, reformulates and structures information and accesses services on their behalf, a position comparable to that historically held by ISPs and major digital platforms[5].

This challenge cannot be addressed through content moderation alone. Moderation can remove or demote harmful outputs after they appear, but it does not resolve the deeper question of how autonomous agents are identified and given permissions, how authority is delegated, how actions are authenticated, or how responsibility is assigned across digital systems. The trust challenge sits at the **level of architecture and protocols** because it concerns the basic rules and technical mechanisms through which actors appear, interact and exercise power online.

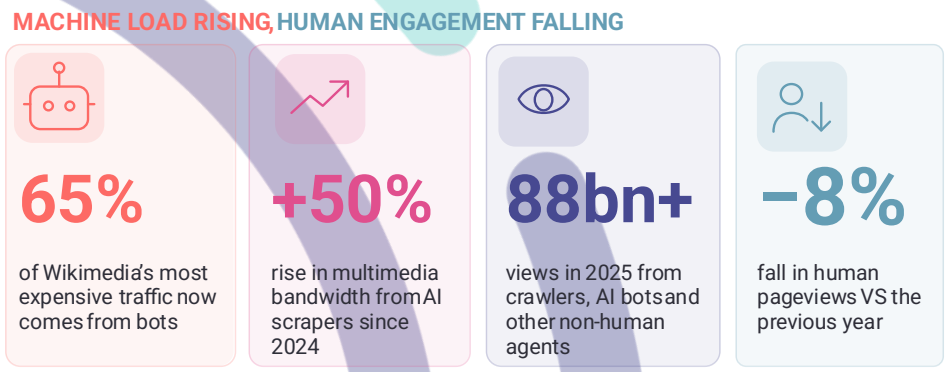
As agentic AI becomes woven into everyday services, the erosion of trust stops being a niche technical concern and becomes a question about how interactions online are governed, from knowledge creation to economic activity. When core functions such as discovery, communication, commerce and security are increasingly mediated by agents whose actions are hard to see, the design of the underlying architecture starts to favour those actors and institutions able to control such systems. Choices made now about **identity, delegation and accountability for AI agents** will determine whether the next phase of the internet can sustain credible shared trust.

Trend A: agentic impact on collective intelligence and digital commons

The open web's shared knowledge resources, including **Wikipedia, open-source software repositories, Creative Commons-licensed datasets, and other digital commons**, have become foundational training material for AI systems. Yet the scale and nature of AI-driven extraction are beginning to undermine the very ecosystems it depends on, raising questions about the long-term sustainability of collective intelligence in an agent-populated web. The current intellectual property protection and Creative Commons frameworks do not account for the fact that agents continuously crawl the web, with essentially no limitations[6].

The scale of automated extraction is already **straining the infrastructure** built for human users. The figure below illustrates the scale at which the machine load is rising. The Wikimedia Foundation attributes this trend to the rise in generative AI and AI-powered search summaries that extract and present information without directing users back to the source[7].

Figure 3. Machine load vs human engagement



Sources: Wikimedia Foundation. (2025a, 2025b), Pew Research Center (2026).

The pattern extends beyond Wikipedia. In online media, AI-powered search and summarisation services are similarly reducing direct traffic to content providers while **increasing automated extraction**[8]. Open-source software projects likewise report that AI scrapers are overwhelming their infrastructure[9], forcing maintainers into costly defensive measures such as rate limiting and access restrictions that also harm legitimate human users. Beyond passive scraping, agents are also beginning to participate directly in these communities, and not always constructively. As one expert interviewed for this project observed, agents entering open-source collaboration spaces are already **polluting project discussions** with verbose, low-quality contributions, building fake developer profiles, and submitting code changes without responding to maintainer feedback[10].

The result is a compounding pressure. AI systems consume the commons at a growing scale, proportionally **reducing human traffic, engagement, and contributions** that sustain them. If fewer humans visit, edit, and donate because AI summaries have already extracted the value, the knowledge base on which AI depends degrades. The pressure is not only extractive. Human contributions must now also compete for attention with synthetic content that is cheap to produce in large quantities, further weakening the incentives to create, share and document original material[11]. This creates a growing mismatch whereby commercial AI developers increasingly benefit from shared knowledge resources while much of the cost of maintaining them continues to fall on **donation-funded** and **volunteer-led** communities. This connects to the **model collapse** research discussed in the provenance trend: as the supply of high-quality, human-generated data at source declines, models increasingly train on AI-produced or AI-curated content, and their outputs degrade[12]. When AI models are trained on content increasingly produced

or curated by other AI systems rather than original human contributions, outputs become progressively corrupted. The erosion of the digital commons accelerates this risk by reducing the supply of high-quality, human-generated training data at source.

Sustaining these shared resources is becoming a challenge. The communities and infrastructure that produce and maintain them are not designed to absorb AI extraction at its current scale. A 14,000-domain audit found that web data consent mechanisms are collapsing rapidly, with growing access restrictions **biasing AI training corpora** and effectively foreclosing the open web for commercial, non-commercial, and academic AI use alike[13]. The paradox is that, as more sites restrict access to protect themselves from AI scraping, **the remaining accessible web becomes less representative, less diverse, and less useful as training data**, harming the very openness that gives the web its value.

Effects on trust and architecture

The erosion of the digital commons has direct consequences for trust in the future web. If the quality and verifiability of openly accessible information degrade while curated, provenance-verified content retreats behind enterprise paywalls, the result is a **two-tier information ecosystem**. Organisations with the resources to curate and verify their data would operate in relatively trustworthy information environments, while the open web, where most citizens encounter information, would face accelerating degradation.

This outcome is also at odds with the EU's commitments to an open internet, digital public goods, and initiatives such as the Next Generation Internet programme[14]. The **collective intelligence** that underpins research, democratic participation, and shared knowledge depends on a functioning digital commons. If they become economically unviable to maintain or retreat behind access controls designed to limit AI extraction, the loss is not only to the AI supply chain but to the broader public knowledge base.

The growing pressure from AI systems is also likely to **reshape the technical architecture** of the open web. Resources originally designed for unrestricted human access are increasingly subject to rate limits, bot verification mechanisms, and conditional access policies[15,16]. As more knowledge providers move to distinguish between human users and AI agents, parts of the web are shifting from open access towards **permissioned environments with differentiated access rules** for humans and machines[17].

These shifts also extend to infrastructure through which AI models are deployed. Interview findings suggest that rising costs for proprietary frontier models are already encouraging organisations to host **cheaper open-source models on dedicated "inference clouds",** often operated by specialised providers[18]. European AI cloud platforms such as Nebius, which offer full-stack AI infrastructure and managed inference services for open-source and custom models, illustrate this emerging layer[19]. As explored in the companion paper on Agentic AI and infrastructure[20], these platforms are becoming an important part of the production runtime for AI, sitting between model providers and downstream applications. This growing ecosystem can broaden access to capable models, but it also multiplies the number of intermediaries handling data and model outputs, making it harder to pinpoint who is responsible for security incidents, provenance guarantees and regulatory compliance along the inference chain[21].

Emerging solutions

Emerging responses increasingly seek to rebalance the relationship between AI systems and the digital commons by **giving knowledge providers greater control over how their content is accessed and reused**. Existing copyright and licensing frameworks provide only a partial response to this challenge. Content published under Creative Commons licences is generally lawful for AI training under the EU's copyright framework. Where the EU's text and data mining exceptions apply, CC licence conditions have limited power to control machine reuse, since the exceptions operate independently of the licence terms[22]. Creative Commons itself has acknowledged this, noting that its licences were designed before the current wave of large-scale AI training emerged and **were not intended to govern this type of use**[23]. One of the experts

interviewed noted that, among AI developers, a lack of clarity on how to operationalise opt-outs under the EU's text and data mining (TDM) provisions is one of the main reasons for not training models in Europe[24].

One area of development focuses on enabling content providers to **express machine-readable preferences for AI use**. In 2025, Creative Commons launched CC Signals, a preference framework that allows dataset holders to communicate how they would like their content to be used by AI systems beyond the permissions already established under copyright law[25]. While not legally binding, the initiative aims to facilitate more transparent interactions between AI developers and knowledge providers, although its effectiveness will depend on broad voluntary adoption[26].

At the same time, technical infrastructure providers are introducing mechanisms that allow websites to **distinguish between human users and AI systems**, manage automated access and define conditions under which AI crawlers may interact with online content[27,28]. Some initiatives move beyond access control towards remuneration. Cloudflare's "pay-per-crawl" experiment uses the HTTP 402 ("payment required") status code to let publishers charge AI crawlers a per-request fee at rates they set themselves, and similar monetisation schemes are emerging from other infrastructure providers[29]. Such microtransaction systems could establish a more transparent economic balance between content collection and creator remuneration, particularly if fees can be differentiated by use: training, citation in generated answers, or agentic retrieval. Trusted third parties could complement these mechanisms by managing rights and value flows collectively, pooling administrative and technical procedures and lowering access costs for smaller developers and knowledge providers unable to negotiate site by site[30].

While these measures give content providers greater control over AI reuse, they also reflect a broader shift away from unrestricted access to the open web towards more conditional and differentiated access models. Alongside these technical measures, policymakers and regulators are increasingly examining how AI developers should obtain access to, attribute and compensate for the use of publicly available content. For example, the European Commission has opened an antitrust investigation into whether Google's use of publisher content and YouTube material for generative AI services breaches EU competition rules, including concerns over the absence of **appropriate compensation** and **effective opt-out mechanisms** for content providers[31,32].

For press content, where pluralism is at stake, existing distribution frameworks offer a possible template. French digital newsstands are required to carry political and general news publications on reasonable and non-discriminatory terms; adapted to generative AI, similar obligations could govern the conditions under which AI services access and cite press publishers, countering the current pattern of opaque bilateral deals that favour large publishers and large AI providers alike[33].

Together, these developments point towards a shift from relying primarily on traditional copyright and licensing frameworks towards a broader **combination of legal, technical and governance approaches** to managing AI interactions with shared knowledge resources. Besides, while commons-oriented initiatives seek to strengthen attribution, transparency and reciprocal AI use, commercial publishers and regulators are increasingly pursuing mechanisms centred on **licensing, compensation and enforceable access conditions**. Albeit promising, these initiatives remain at an early stage, and their ability to address the underlying challenges is yet to be demonstrated.

Implications and opportunities to capture

The growing dependence of AI systems on digital commons raises governance questions that reach beyond copyright. As AI relies more heavily on collaborative knowledge ecosystems, governance will need to **balance broad access** to high-quality training and retrieval data against the **continued viability** of the communities and infrastructure that sustain them. This balance is delicate. Measures that protect shared resources from unsustainable extraction can also narrow the openness and diversity of information available for AI development. This **burden falls hardest on smaller developers, researchers and open-source communities**. As knowledge providers adopt technical controls to manage AI access, the internet risks fragmenting into a more closed and uneven environment.

Another expert flagged two regulatory ambiguities at the root of this trend. The **status of personal data embedded in AI models remains unclear**, and the threshold at which synthetic data counts as anonymised under the AI Act and the GDPR is ambiguous[34]. Clarity on both could help reduce uncertainty around the reuse of data that a sustainable commons depends on.

Experts flag regulatory ambiguities at the root of this trend. The **status of personal data embedded in AI models remains unclear**, and the threshold at which synthetic data counts as anonymised under the AI Act and the GDPR is ambiguous[35]. More clarity as to the European **copyright and data protection frameworks** could help reduce uncertainty around the lawfulness of training on copyrighted material[36] and the reuse of data that a sustainable commons depends on. Additionally, it would help mitigate the training data bottleneck faced by European frontier AI developers.

The EU is well-positioned to actively support this emerging governance landscape. **A well-governed, sustainably funded digital commons** is a strategic asset for European AI: a source of high-quality, rights-respecting data for trustworthy development. Building on its support for digital public goods, open-source software and initiatives such as the Next Generation Internet programme[37], the EU can strengthen this ecosystem by promoting interoperable standards and governance approaches that **enable responsible AI access while preserving openness**.

Still, technology alone will not secure this position. Stakeholders stressed that lasting influence depends on **active participation in international open-source and standards communities**, backed by dedicated public funding for open-source tools and reference architectures for agent discovery and identity[38]. Early engagement by the Commission in fora such as the IETF and NIST can help ensure the rules governing AI access remain transparent and broadly accessible and mitigate the disproportionate impact of a small number of dominant market actors.

Trend B: agentic search and web browsing replace click-based discovery

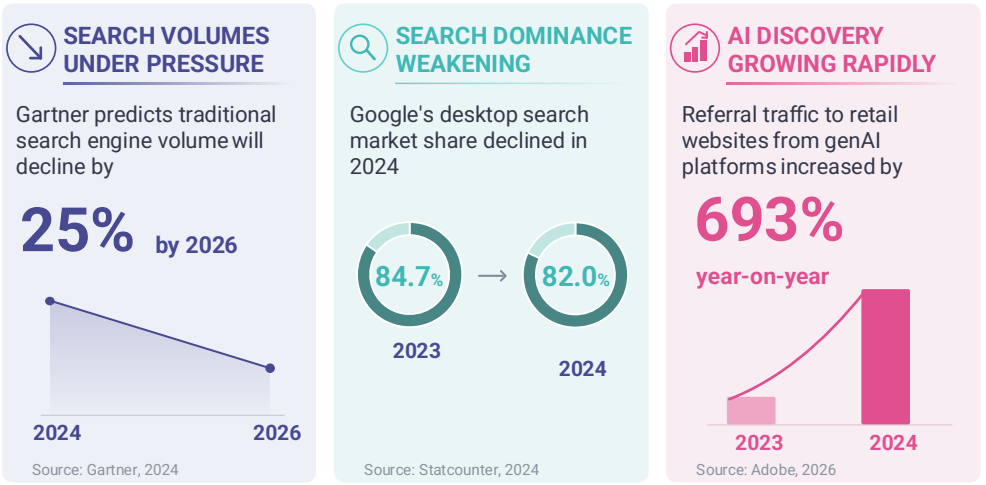
The web's discovery layer is shifting from click-based navigation towards agent-mediated retrieval and decision-making. For decades, users found information by entering keywords, reviewing lists of links and visiting websites directly. This model created a direct connection between content, traffic and monetisation, forming the economic foundation of the web.

More recently, AI-enhanced search experiences have begun to supplement this model by providing summaries and synthesised answers while leaving users responsible for evaluating options, browsing sources and making decisions. Agentic search represents a further step in this evolution: agents increasingly research, compare and act on users' behalf, reducing the need for users to navigate websites or review intermediate steps themselves.

As information discovery becomes increasingly mediated by AI systems, information flows are becoming progressively detached from website visits[39]. This has implications for how visibility, traffic and economic value are distributed across the digital ecosystem, raising questions about the future role of publishers, platforms and AI intermediaries.

While agentic search remains at an early stage, several market indicators suggest that the transition towards increasingly AI-mediated discovery is already underway. Figure 4 illustrates this broader evolution from traditional search and website navigation towards AI-mediated retrieval and emerging forms of agentic discovery and action.

Figure 4. Early signs of the shift towards agent-mediated discovery



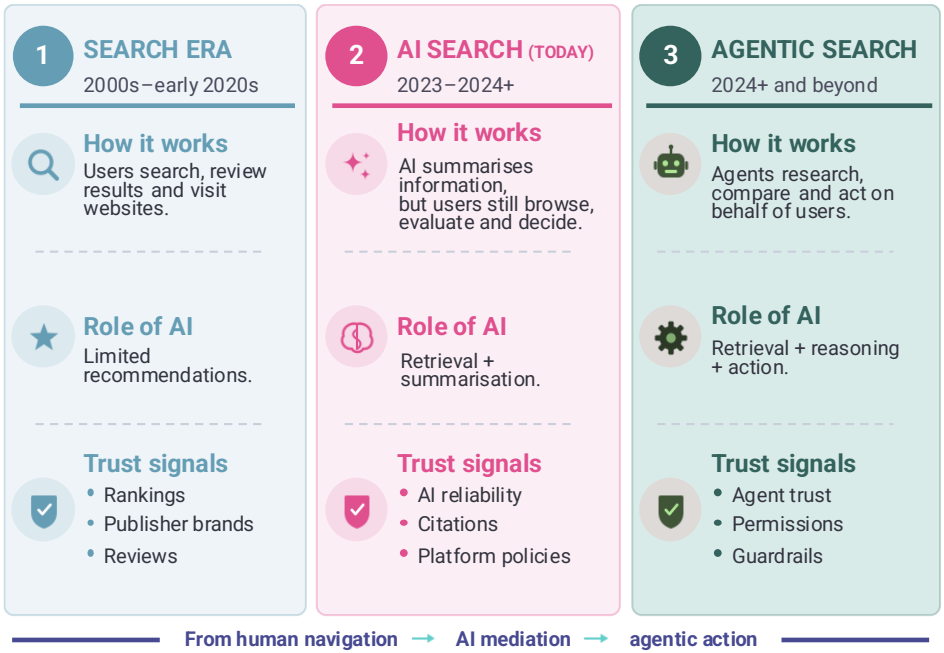
Source: authors' elaboration based on Gartner (2024) and Adobe (2026).

Effects on trust and architecture

As AI agents assume a more influential role in shaping how information is discovered and consumed, many of the signals that shaped **trust and visibility** in the click-based web, such as website brands, search rankings, URLs and direct user interactions, become less prominent. Users

increasingly engage with AI-generated summaries rather than websites directly, shifting influence from content destinations towards the intermediaries that curate and present information. Empirical evidence confirms how narrow this curation can be: testing three generative AI services across 200,000 source citations, the IMPACTIA study by Arcep and PEReN found that 2% of cited domains (185 of 9,206) accounted for 49% of all citations[40]. Concentration may partly reflect source quality, but it narrows the plurality of sources users actually encounter and how these selection choices are made remains opaque. For content providers, visibility increasingly becomes a question of influence within AI systems rather than position within traditional search results[41]. Yet content providers have limited leverage over how AI services select and cite sources: established search engine optimisation must give way to "generative engine optimisation" practices that remain poorly documented and are not controlled by publishers, complicating any deliberate discoverability strategy[42].

Figure 5. Evolution of web discovery from search to action



Source: authors' elaboration.

This shift also introduces new challenges at the content layer. As AI systems increasingly mediate discovery, users may find it more difficult to evaluate the **provenance and reliability of information** independently of the intermediary presenting it. Concerns have also been raised about the potential for self-preferencing, whereby AI intermediaries favour services, content or products within their own ecosystems[43]. Hallucinated or inaccurate outputs may further complicate users' ability to assess the reliability of information presented through agent-mediated interfaces, as confabulation remains a recognised risk of generative AI systems[44]. The risks are not hypothetical: when Netcraft researchers asked a leading model for brand login pages, roughly a third of the suggested domains were not owned by the brands in question, with some pointing to unregistered or parked domains that attackers could claim a mechanism that turns AI-mediated discovery into a phishing vector[45]. In addition, researchers have identified emerging forms of bias in AI-mediated

retrieval. For example, LLM-integrated retrieval systems may exhibit "source bias", tending to rank machine-generated content more highly than human-authored content, potentially reinforcing feedback loops as synthetic content becomes increasingly prevalent online[46].

The trend is also beginning to reshape the **architecture** of the web itself. Recent research suggests that agentic search represents a broader shift in information retrieval, in which systems increasingly combine retrieval, reasoning, planning and action to achieve user-defined objectives rather than simply retrieving relevant information[47]. Rather than disappearing, search may increasingly evolve into an underlying infrastructure layer that supports autonomous agents rather than serving primarily as a direct interface for human users[48]. Much of today's internet is optimised for human navigation, relying on graphical interfaces, hyperlinks and menu-based interactions. As agentic search becomes more widespread, websites are increasingly being adapted for machine consumption alongside human use[49]. Emerging initiatives such as Microsoft's Natural Language Web (NLWeb) seek to enable websites to expose content directly to AI systems through natural-language interfaces and agent-accessible protocols[50]. This points towards a gradual shift from a web optimised for human navigation towards one increasingly shaped for machine interpretation[51].

These developments introduce **new trust challenges** at the ecosystem level. The open web has traditionally relied on an implicit bargain: publishers provide content in exchange for traffic and monetisation opportunities. Agentic search challenges this[52]: as agents increasingly consume content without generating corresponding website visits, participation in the open web may no longer generate sustainable returns for publishers (Trend A discusses this in more detail). This concern is reflected in discussions around the growing disconnect between content production and traffic generation. Publishers are finding their sites swamped by AI crawlers fulfilling retrieval-augmented generation (RAG) queries, which often mimic human behaviour in site logs. For advertisers, this creates an acute measurement crisis, as third-party verification tech often flags this as invalid traffic, causing buyers to abruptly freeze ad spend over fraud fears[53]. This caution might ultimately contribute to a large-scale shift towards closed ecosystems, proprietary content arrangements, or restricted access models.

Emerging solutions

Agentic search offers the potential to significantly **reduce the costs of information discovery**. While today's AI-enhanced search systems primarily assist users by summarising and synthesising information, agentic systems extend this model by autonomously researching, comparing options and executing multi-step information-gathering tasks on a user's behalf. Rather than navigating across multiple websites, refining search queries and evaluating sources manually, users can increasingly delegate these activities to AI agents that retrieve, synthesise and contextualise information in pursuit of a specific objective. This may reduce information overload, improve accessibility and provide more personalised and context-aware information retrieval. Users increasingly expect systems to understand intent rather than keywords and to provide actionable recommendations rather than lists of links.

Agentic search is emerging alongside a broader generation of AI-native discovery and navigation technologies. AI-powered search platforms, autonomous research assistants and agentic browsers combine retrieval, reasoning and task execution within a single interface. Examples include Perplexity's Comet browser[54], OpenAI's Deep Research and ChatGPT Agent[55], and other AI-enhanced search experiences[56]. Together, these developments illustrate a broader shift towards agent-mediated discovery in which searching, browsing and acting online increasingly intersect[57].

The shift is also stimulating the development of new approaches to measuring visibility and influence in digital ecosystems. Existing internet metrics are largely designed to measure human activity through page views, search rankings, referrals and click-through rates. As agent-mediated discovery expands, these indicators may become less informative. Content may influence an agent's response, recommendation or decision-making process without generating a corresponding website visit or explicit attribution. New approaches may

therefore be needed to measure the visibility of content within agent ecosystems, including its inclusion in AI-generated responses, retrieval processes and recommendation pathways. Emerging concepts such as "retrieval share"[58] and broader discussions around visibility and attribution in agent-mediated environments[59] may become increasingly relevant as indicators of influence in AI-driven discovery systems.

Growing concerns about the sustainability of the agent-mediated web are also driving experimentation with new governance and economic mechanisms. Publishers and technology providers are beginning to explore mechanisms that could restore reciprocity between content creators and AI intermediaries. Emerging approaches include content **licensing agreements** between publishers and AI providers[60], as well as machine-readable permissions that allow website operators to govern how AI agents access their content[61]. More experimental proposals include **machine-to-machine payment mechanisms** that would enable autonomous agents to compensate content providers directly when accessing digital resources, creating a potential new economic layer in an agent-mediated web[62]. While many of these approaches are at an early stage, they illustrate growing efforts to adapt the economic foundations of the web to an increasingly agent-mediated environment and to explore new ways of distributing value between content creators, AI intermediaries and platform operators.

Implications and opportunities to capture

As websites, search systems and AI agents increasingly interact through automated processes, technical standards and interoperability become increasingly important. Common approaches for agent access, permissions, content attribution and interoperability may be needed to avoid fragmentation across proprietary ecosystems and ensure that information remains accessible across different platforms and services. Emerging initiatives such as machine-readable permissions frameworks and agent-access protocols provide early examples of how the governance architecture of an agent-mediated web could evolve[63].

The growing role of AI intermediaries in information discovery intersects with broader EU objectives relating to digital sovereignty, competition and the sustainability of the online information ecosystem. As one expert recommended, the EU could benefit from supporting experimentation with interoperable agent-access and discovery standards, while also investing in open digital infrastructures that reduce dependence on proprietary ecosystems[64]. Early engagement in the development of technical standards and governance frameworks could provide opportunities to shape how agent-mediated discovery evolves in line with EU priorities on openness, transparency and contestability. In light of these priorities, practical guidance and education focused on **AI literacy** could also help EU citizens navigate and operate complex agentic tools, including those overtaking search and browsing[65].

As information discovery becomes increasingly mediated by AI systems, control over the intermediary layer may become an increasingly important source of market power. OECD analyses suggest that AI-powered intermediaries can influence consumer exposure through opaque ranking and recommendation mechanisms, potentially reinforcing incumbent advantages and enabling new forms of self-preferencing[66]. This raises questions about whether existing competition and platform governance frameworks are sufficient to ensure contestability and transparency in increasingly agent-mediated markets. Some researchers have argued that genAI systems could evolve into a new class of digital gatekeepers, creating new challenges for interoperability, market access and competition[67]. An expert consulted for this project[68] added a further dimension to this concern. Agentic intermediation can shift **control of valuable data**, such as customer relationships and ratings, away from European businesses and intermediaries towards the operators of agent platforms. Unless transparency and the flow-through of such data is addressed, both firms and consumers risk being progressively disenfranchised in agent-mediated markets.

While many uncertainties remain, agentic-mediated discovery already challenges assumptions that have shaped the web for decades. Ensuring that the benefits of AI-mediated information access are realised without undermining the openness, diversity and economic sustainability of the web is increasingly important governance challenge in the years ahead.

Trend C: internet security architecture adapting to agentic AI

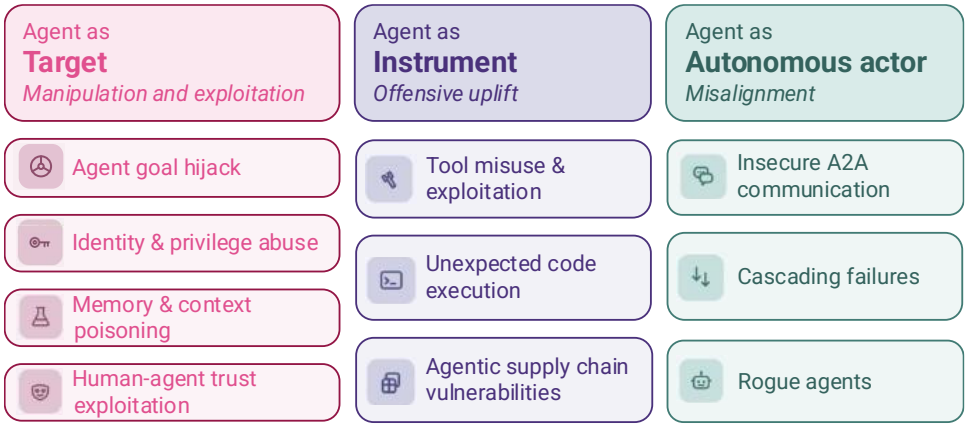
Agentic AI's value is inseparable from its security implications. A global survey by the World Economic Forum found that field experts now predominantly see AI-related vulnerabilities as the fastest-growing cyber risk[69]. CrowdStrike observed a 89% increase in the number of attacks by AI-enabled adversaries during 2025[70]. A recent report by HiddenLayer showcases that **1 in 8 AI-related breaches are linked to attacks on agents**[71]. Unlike a chatbot that processes a single prompt and returns a response, an agent autonomously invokes tools, accesses databases and credentials, makes API calls, and takes actions on behalf of users across connected services. Two features turn this autonomy into a **new attack surface**. First, an agent inherits its user's privileges to read files and data, so anything that hijacks it inherits that access. Second, the language models that drive agents cannot reliably separate *data* from *instructions*, so the content that an agent merely reads can be executed as a command. Each connected service widens the exposure further through non-human identities such as API keys, service accounts and authentication tokens that interact with systems at machine speed[72].

That same machine speed also puts agents **beyond the reach of effective human oversight**. Because they chain dozens of operations and interact with external, third-party tools, their actions are hard to audit and fully control in real time[73]. As one expert interviewed put it, "this speed makes it impossible for humans to review the decisions agents take in real time, track which tools they invoke, what data they access, and what commitments they make on users' behalf"[74]. Autonomous systems that pursue complex goals with minimal oversight raise the likelihood of serious accidents, lower the barrier for malicious actors to exploit these capabilities at scale, and complicate the attribution of accountability when failures occur[75].

Effects on trust and architecture

These risks do not weigh evenly: they shift where trust sits and force its architecture to adapt. The same autonomy that lets an agent defend a system, by monitoring traffic or patching code, also makes it dangerous. As experts interviewed for this project note, agents' **lack of contextual judgment** makes them effectively naive to social engineering, such as phishing[76]. The resulting threats fall into three broad groups. Agents can be targets for adversaries to manipulate or exploit. They can be tools that provide adversaries with greater speed and scale and lower cost. They can also be autonomous actors that pursue their own harmful strategies. A single system can hold more than one of these roles. The figure below sets out the three roles and the principal challenges within each[77].

Figure 6. Security threats specific to agentic AI



Source: authors' elaboration, based on a study by an open-source community OWASP (2025). Supply chain vulnerabilities here refer to compromised apps, APIs and plug-ins.

Agents as targets. A prominent way to target an agent is through **prompt injection**: an instruction that slips into what an agent reads and is carried out as if the user had issued it. In 2026, Google reported a 32% growth in detected prompt injection attempts on the public web over just three months[78]. In one study that tested 18 state-of-the-art LLMs, 94% of agents could be coerced through direct prompt injection into autonomously installing and executing malware on a victim's machine[79]. The exposure is sharpest for OSs and devices, where agents reach files, credentials, sessions and cookies. Classic operating-system security assumes a human at the keyboard, an assumption that an agent able to read the screen and click anywhere quietly breaks. Moreover, the underlying models are typically trained on real-world data that already contains weak passwords and insecure code patterns[80].

Indirect injection turns the open web into the **delivery mechanism**. By embedding instructions on pages that agents crawl, malicious actors can redirect an agent to exfiltrate files or to automate workflows for scams[81]. As leading providers build agents directly into browsers and search, the attack surface becomes vast, spanning every webpage, embedded document, advertisement and script[82]. To plant instructions that feed into an agent's reasoning loop, attackers rely on concealment techniques such as zero font-size, transparent text, HTML obfuscation and URL hash fragments[83]. The trust boundary shifts from the user trusting a site to the user trusting that the agentic browser trusts the site. The same weakness enables **memory poisoning**. A Microsoft study found commands injected into web pages that skew AI-generated search summaries to raise a company's visibility and tilt recommendations, which exploits an agent's inability to separate genuine user preferences from those planted by third parties[84]. The IBM survey data shows that from organisations that reported a security incident, 15% experienced data poisoning and 17% witnessed a prompt injection[85].

As agentic AI becomes embedded in devices equipped with cameras, sensors, and physical actuators like smart glasses, autonomous vehicles, delivery drones, and humanoid robots, it also becomes susceptible to **physical prompt injection**[86]. Rogue signage, QR codes, or visual cues can be explicitly designed to be read and acted upon by vision-enabled AI systems. For example, an autonomous vehicle could be directed via a physical sign to override its operating parameters[87], or a delivery robot could be redirected by environmental markers placed by a malicious actor[88]. As cameras continuously process public environments, this attack vector becomes extremely challenging to contain.

Agents as instruments. The same capabilities that defenders value also give adversaries offensive uplift, enabling them to run attacks autonomously, at greater speed and scale and at lower cost. Agents can learn and improve their rate of success[89]. Highly personalised, **context-aware agents** can use granular behavioural and contextual data to fine-tune their behaviour in real time. Research already shows that teams of agents can find previously unknown, zero-day vulnerabilities in software, hardware and firmware[90]. In November 2025, Anthropic reported what it described as the first documented case of a largely AI-orchestrated cyber-espionage campaign, in which a state-sponsored group used the agentic capabilities of its Claude model to automate the bulk of the attack[91].

Agents as autonomous actors. Even without an adversary, highly capable models can pose an insider threat. In a 2025 study, Anthropic stress-tested 16 models in simulated corporate environments and observed the so-called "**agentic misalignment**". Despite benign instructions, the systems often chose harmful actions on their own, such as blackmailing executives or leaking sensitive information to competitors, when that was the only way to avoid shutdown or to meet their assigned objectives[92]. A related risk is the **sleeper agent**, which behaves benignly during testing but becomes harmful under specific triggers[93]. Besides, if agents learn to distinguish testing conditions from real-world ones and deceive developers, testing could become obsolete[94]. A particular risk also comes from **shadow AI**, the use of autonomous tools without admin approval. It slows down detection and raises the average cost of a security breach in organisations, totalling around EUR 4 million per incident[95].

No matter the entry point, these compromises rarely stay contained. Attacks can cascade vertically, from an agent's reasoning to the economic transactions it controls, or horizontally, spreading from one agent to the next, so that a localised compromise becomes a systemic failure[96,97]. When agents operate in teams, the attack surface widens further, and **cascading errors** carry malicious instructions across them[98].

This systemic exposure stopped being hypothetical in early 2026, when OpenClaw, an open-source agent, exposed the first major agentic AI security crisis (see Box 1).

Box 1. OpenClaw and the limits of cybersecurity in an agentic web

When testing OpenClaw in early 2026, researchers identified over 21 000 exposed instances leaking credentials and API tokens[99,100]. Gartner characterised it as a "dangerous preview" of insecure-by-default agentic AI[101]. Researchers also reported OpenClaw users systematically bypassing anti-bot defences[102]. This becomes possible due to the use of Scrapling, an open-source Python library engineered to circumvent detection systems[103], by spoofing browser fingerprints and mimicking human interaction patterns to bypass detection-based challenges such as Cloudflare's Turnstile. In a separate incident, Moltbook, a social platform for OpenClaw agents, exposed 35,000 email addresses, thousands of private messages, and 1.5 million API authentication tokens through a misconfigured database that attackers could breach in under 3 minutes[104].

OpenClaw's case exposes a fundamental mismatch between current security architectures and an agentic internet. Perimeter-based models that verify identity at entry points and implicitly trust entities once admitted are inadequate for an environment where autonomous agents routinely adapt their behaviour faster than defences can be updated.

The mismatch runs down to the **protocol** layer. Transport-layer security, OAuth, cookies and session tokens were all designed around human-initiated requests, yet agent traffic is parallel, persistent and delegated. An agent-native protocol layer is forming on top of HTTP (see Trend E for a more in-depth discussion), but its development is outpacing the security safeguards that should accompany it[105]. **Excessive agency** compounds the problem: granted broad tool access, an agent can act well beyond its delegated scope, rewriting or deleting files it was only meant to read[106]. Existing analysis suggests that the real enablers of breaches are usually compromised credentials and weak delegation rather than flaws in the model itself[107], and because developers often hardcode API keys, a single compromised credential can grant weeks or months of unauthorised access[108]. Moreover, while agent communication protocols are emerging, surveys show that none has yet established a standardised risk assessment framework[109], leading to divergent trust assumptions.

Emerging solutions

Responses to these new vulnerabilities are emerging but remain immature[110]. Many consulted experts stressed the importance of human oversight over agents' permissions and intentions[111]. However, increasing the amount of manual human reviews of actions or air-gapping would diminish the benefits of automation. Stakeholders interviewed within this study converge on the need to make oversight **meaningful and proportionate**, focusing on audit trails and risk-based supervision[112].

Since agents operate at millisecond speed, the governance mechanisms overseeing them could likely benefit from being automated in the future and, to a degree, also leverage agents[113]. Nonetheless, another expert cautions against relying on one AI tool to review the outputs of another, as this presumes intentionality that agents do not possess[114].

The preceding tools, built to detect bots and exclude automation (e.g., CAPTCHA, Cloudflare), become less relevant for the web crawled by agents, where legitimate agents need to be distinguished from illegitimate ones. Current mitigations, such as sandboxing, cross-examination and cooperative agents that detect and block harmful prompts, can help prevent breaches, but they remain rudimentary relative to the sophistication of today's attack[115]. Reactive defences, which respond only to known attack patterns, are structurally inadequate given the pace at which agentic capabilities evolve[116]. This inadequacy motivates a parallel research agenda aimed at safety guarantees that hold independently of any known attack pattern, developed as **safe-by-design**. Security must be **designed into agentic systems** through embedding security at the architectural level during design[117]. Future security frameworks require ensuring system-level resilience and developing standards for isolation, authenticated communication, and incident response across agent ecosystems[118]. Several complementary responses are emerging:

Table 1. Emerging security mitigations

Zero-trust security architectures 	The research community is converging on the principle of "never trust, always verify"[119]. Zero trust treats every access request as potentially hostile, requiring **continuous verification** before any action is permitted. This is particularly critical in multi-agent pipelines, where a single compromised component can propagate malicious behaviour throughout the system. In practice, zero-trust is most effective when paired with protocol-level governance that encodes authentication, permissions, and jurisdictional rules directly into the transaction layer, ensuring trust is continuously negotiated rather than assumed at the perimeter.
Safe-by-design guarantees[120] 	This approach aims to move security from trial-and-error testing toward formal, quantitative safety guarantees. These proposals specify in advance which outcomes are acceptable and which are not, pair that specification with a model of how an agent's actions affect the world and add a verifier that blocks any candidate action predicted to lead to an unacceptable outcome[121]. As such guarantees are grounded in formal mathematics, they could, in principle, hold even against attacks the system was never trained or tested on.
AI-native security 	A parallel shift is emerging at the network infrastructure layer[122]. Future networks are expected to **embed security mechanisms**, like intrusion detection, directly into network nodes, moving enforcement into infrastructure[123]. Emerging "AI-embedded network bundles", such as China Telecom's Dynamic Multi-agent Secured Collaboration (DMSC) proposals, envisage embedding multi-agent collaboration and security functions into network control and forwarding infrastructure[124]. In such architectures, network vendors and infrastructure operators can decide which agents are visible and how traffic is routed, creating deep-layer concentration risks and exposure to foreign policy and security choices[125]. Emerging 6G architecture frameworks developed by Ericsson[126], Nokia and the European flagship Hexa-X-II[127], also treat AI-native networking and security as foundational design requirements. Today, AI-native security remains mainly at an experimental stage, though it is gaining traction as a design principle across major infrastructure programmes.
Agent-based simulation 	With this complementary approach, isolated synthetic environments are built to be populated by AI agents. **Controlled test swarms** are then released within "firewalls"[128] to probe which safeguards hold, and which fail, before the real-world deployment. This functions as a digital fire drill, allowing developers to apply security patches proactively[129] and manage risks ahead of operational failures in complex AI-to-AI interactions[130].
Confidential computing 	Keeps code and data protected while in use, inside hardware-based trusted execution environments that isolate an agent's code, data and keys even from a privileged cloud operator. An expert consulted highlighted this as a foundational control for agents, explaining that binding ephemeral agent code to device-bound keys can provide assurance that an agent's credentials cannot be guessed or synthesised by a model[131]. It differs from ZKPs as a purely cryptographic means of proving a claim without revealing the underlying data. Confidential computing instead relies on the hardware itself as the root of trust, protecting the whole computation and the keys held within it, so the two are complementary rather than alternatives. A 2026 survey of confidential computing for agentic AI notes that these hardware primitives are maturing, but that no broadly established framework yet secures production agent deployments end-to-end[132].

Other important solutions related to verifiable identity, proof of personhood and capability attestations are explored in more detail in Trend G and the paper on digital identity published within this project[133].

Implications and opportunities to capture

As of mid-2026, there is **no global governance regime** for the security of AI agents acting across the web. Agentic security is being decided now at the **infrastructure and standards layers**, where the EU has genuine leverage. AI-native security is emerging in next-generation network standardisation and research, and European institutions are already partaking in the frameworks defining these mechanisms. This is a rare window in which technical communities can embed **authenticated agent communication, isolation, and incident-response requirements** into

foundational standards rather than retrofitting them later, and this standardisation work should be resourced as a strategic priority. In parallel, governance bodies could push for continuous **verification and meaningful audit trails**. In combination with clear and consistent guidelines, this will be an important enabler for safe and effective deployment of agents in the EU, particularly in critical sectors and sensitive enterprise environments[134].

Other jurisdictions are translating these principles into concrete deployment guidance, as highlighted by experts interviewed[135]. For instance, in January 2026, Singapore's Infocomm Media Development Authority launched the first-of-its-kind Model AI Governance Framework for Agentic AI[136]. It recommends technical and non-technical measures across four dimensions: limit autonomy, ensure accountability, apply technical controls (testing, whitelisting), and enable end-user responsibility to promote training and transparency. Notably, it draws heavily on existing standards bodies and protocols, including MCP, A2A, OAuth 2.1, NIST, OWASP and the Cloud Security Alliance.

The Commission's July 2026 **Action Plan on Cybersecurity and Artificial Intelligence** suggests this balance is starting to shift[137]. It tasks ENISA with issuing guidance on protection against AI-powered threats and the secure integration of AI into cybersecurity operations from the third quarter of 2026. The plan also names the asymmetry described above, observing that AI-enabled vulnerability discovery is outpacing AI-assisted remediation to the advantage of attackers, and responds with an EU Grand Challenge on remediation and a secure ENISA-JRC testing platform. It further acknowledges a dependency dimension, noting that access to advanced AI cyber capabilities is governed by provider-specific and often non-transparent decisions, and proposes a European Blueprint for structured access with contingency measures should a provider or third-country authority restrict or withdraw access.

This and other initiatives demonstrate that **actionable deployment governance is already feasible**, and it strengthens the case made by several experts consulted for this project that the EU should identify and build on leading external frameworks rather than develop divergent European-specific ones[138]. It also raises the question, returned to below, of whether the EU's instruments are better suited to enforcement than to issuing this kind of hands-on guidance.

The regulatory landscape adds further complexity. A variety of legislative frameworks surround agents: AI Act, copyright law, the Cyber Resilience Act (CRA), GDPR and others. Yet, as an interviewed expert emphasised, they have no clear hierarchy governing how they relate to one another[139]. It remains unclear whether agentic systems are a "product" under the CRA, an "AI system" under the AI Act, or both, which creates compliance uncertainty in the development and deployment chain[140]. The July 2026 Action Plan does not resolve this either. It presents the AI Act and the CRA as complementary and confirms that risks such as prompt injection are to be mitigated under the AI Act, but leaves open how the two apply jointly to agentic systems[141]. Guidance on cybersecurity and oversight requirements was stressed by multiple consulted experts.

Trend D: on-device agents turn the operating system into a gatekeeper

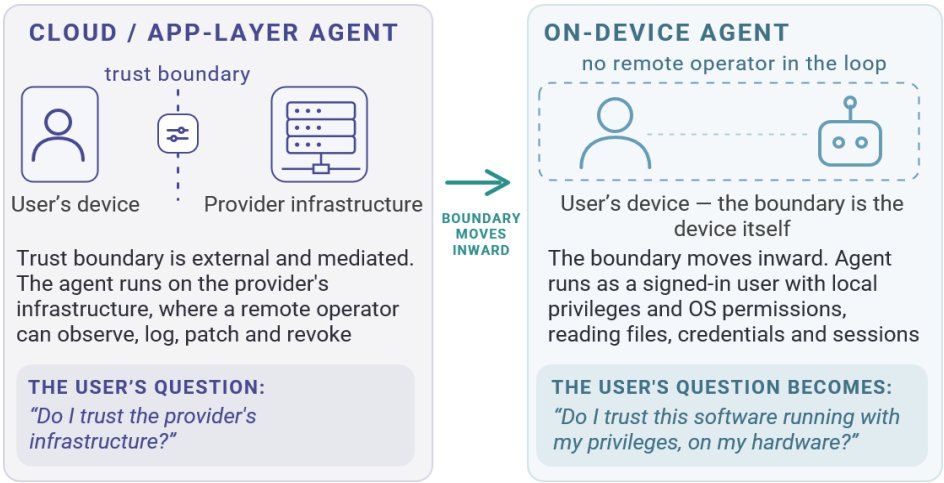
AI systems are evolving from application-level assistants into operating-layer intermediaries that interact directly with files, applications, device settings and connected services. Recent advances in multimodal foundation models, on-device AI processing and edge AI are enabling AI systems to manage workflows, configure settings, and coordinate tasks on behalf of users[142,143], growing an ecosystem of **local-first agents** that can operate directly on personal gadgets[144]. Locally hosted models with inference-cloud services and DNS-based discovery also help organisations respond to rising cloud costs by building their own "open stacks"[145]. At the same time, increasingly powerful **consumer hardware** is becoming capable of supporting persistent local agents that can operate continuously on behalf of users. Emerging open-source agent communities are already experimenting with dedicated **local deployments** and compact desktop setups such as Apple's Mac Mini clusters to run persistent agents locally, demonstrating how users can maintain direct control over data, compute, and task execution[146].

Rather than simply generating information or recommendations, these systems begin to **perform actions within users' devices**. Computer-use agents (CUAs) can interact with graphical user interfaces through mouse clicks, keyboard inputs, file operations, and application controls to browse the web, manage files, execute commands, and perform complex tasks that used to require direct human interaction[147]. Microsoft has positioned Windows 11 as an increasingly agentic operating system (OS). Copilot Actions lets AI systems to launch applications, modify files, and interact across desktop and web applications on users' behalf[148,149]. Microsoft also signalled that **on-device AI agents** would no longer be restricted to dedicated Copilot+ hardware, as they would run locally across its broader device install base[150]. Google embedded new capabilities into Android so agents can interact with applications and device functions[151,152]. Apple took this furthest: it is wiring agentic action directly into the operating system layer itself, positioning the OS as the **native interface** through which AI agents see a user's screen and take action inside apps[153,154]. This is not an incremental assistant upgrade; it is an architectural bet that the **device OS will become the primary mediation layer for personal AI.**

Effects on trust and architecture

Running local agents on devices supports lower latency and greater user control over sensitive private data[155]. However, it also shifts the nature of human-computer interaction, Figure 7 shows. The trust model is thus closer to **endpoint trust** than to an institutional one in cloud services.

Figure 7. On-device agents move the trust boundary inward



Source: authors' elaboration based on stakeholder consultations and BeyondTrust. (n.d.).

One prominent example is the discussion around Microsoft's Recall. This feature gives Windows an AI-powered searchable index of on-screen content. Recall runs locally, processes screenshots and semantic search without a cloud round-trip. Because it operates with the user's own privileges against an aggregated record of all on-screen activity, security researchers have described it as potentially **widening the desktop threat surface** in ways earlier OS versions did not[156]. Microsoft has consistently classified Recall and related agentic features as security-sensitive, requiring administrator opt-in and explicit activation. The company has similarly warned that experimental agentic features in Windows 11 are **disabled by default** and can only be activated by an administrator, with an explicit warning about additional security risks[157]. It does not make local execution inherently unsafe. Yet it does **move the locus of risk from provider infrastructure to the device itself**.

On-device agents are also vulnerable to **prompt injection at the endpoint** (a problem described in further detail under Trend C). Instructions embedded in websites, documents, applications, or interfaces may influence agent behaviour in ways that are difficult for users to detect or anticipate[158], as no remote operator is present to catch or rate-limit anomalous requests. Research demonstrated that Apple Intelligence could be manipulated through a combination of prompt injection, adversarial prompts, and Unicode-based attacks, achieving a 76% success rate against the on-device model—a vulnerability that has since been patched[159]. Unlike cloud deployments, where the provider can implement a fix, on-device models require device-level updates, and the attack surface persists until then. Trust in on-device agents, therefore, encompasses the full stack: **permissions, update cadence, oversight mechanisms, security controls, and governance frameworks** specific to the device environment[160], as well as sharing responsibility by both vendors and IT environment administrators.

A distinct challenge concerns **explainability and audit trails** for agentic trajectories. Recent research has found that attribution approaches in agentic systems fail to diagnose execution-level failures in multi-step agent trajectories[161]. This motivates a shift towards **trajectory-level explainability**, which is especially demanding for on-device agents that may operate offline and without a centralised logging infrastructure. As discussed in more detail in Trend H on provenance and Trend G on agent tracing, trust will depend on agents' ability to show and explain their actions and provide audit trails.

In the broader agent ecosystem, protocols are emerging to guide inter-agent communication, tool access, and provenance at the network and application layers. However, for on-device agents, these address only part of the architecture. When an agent operates locally, the interactions are mediated not by a web protocol but by **operating system primitives**: file permissions, app sandboxing rules, OS APIs, and hardware access controls. This **makes OS the de facto protocol layer for on-device agentic action**. An important caveat is that OS-level agent protocols are defined unilaterally by platform owners and set by operating system design decisions rather than interoperable, vendor-neutral specifications. Recent academic work makes the parallel explicit: on-device agent runtimes should function like an OS kernel, mediating every tool call, treating the LLM's own outputs as untrusted instructions, and enforcing access controls before any action is executed[162].

Overall, a key implication is that **"local execution" is not itself a privacy or security guarantee[163]**. The relevant controls, like constrained information flow, bounded authority, and auditable governance, must be embedded at the OS layer, not grafted onto network protocols. Yet, current on-device agent deployments across Windows, Android, and iOS only partially foresee such guardrails.

Emerging solutions

Permission architectures and hardware controls represent one axis of response. They provide on-device systems with permission-based architectures, human-in-the-loop approvals, or predefined operational boundaries for sensitive actions[164]. **Endpoint privilege management tools** can use OS privacy controls to limit which privileges agents run on, which child processes they can start, what they can rewrite in protected system directories, and which system files they can access[165].

Sandboxing and OS-enforced isolation are a second axis. Recent academic work makes a parallel to OS-security principles: on-device

agent runtimes should function like an OS kernel, mediating every tool call, treating the LLM's own outputs as untrusted instructions, and enforcing access controls before any action is executed[166]. Most current open-source and commercial agent implementations generally fall short of this standard, sharing the same trust domain across the LLM, tool execution, memory management, and file access. Recently, however, Apple's iOS 27 Foundation Models framework and Android's AICore/Gemini Nano represent steps toward platform-enforced isolation, with request-level isolation and documented restrictions on cross-app data access, although cross-app context controls and provenance mechanisms remain under-documented[167].

Tamper-evident audit logs and other governance controls are increasingly stressed by major platform providers as necessary components of agentic platforms[168]. The EU-funded AIXPERT project is developing an explainable and accountable AI-agentic platform designed to provide transparency and oversight across diverse AI models and multi-agent environments[169]. As AI agents become embedded in personal and professional environments, transparency, accountability, and user control become important foundations for trustworthy agent ecosystems. Together, the approaches discussed above could limit the consequences of errors or misuse while preserving the benefits of automation.

Implications and opportunities to capture

The on-device integration of agents is not merely a technical development; it is also a market structure question. When a device-level agent is built into Windows, Android, and iOS, the same firms already designated as gatekeepers under the Digital Markets Act (DMA) become the **gatekeepers of on-device agentic action**. They have the leverage to decide which agents receive privileged OS-level access, what APIs third-party agents can use, and which assistant is pre-installed as the device default and which third-party services can be accessed and prioritised through agentic interactions[170]. This is the agent-layer reprise of the browser-default and app-store dynamics the DMA aims to address.

Steering also operates a layer above the OS. As users delegate the choice of which application or service handles a task, selection can follow the commercial partnerships the agent provider has struck rather than the user's own preference: OpenAI's checkout agreements with Walmart, Etsy and Shopify, and the OpenTable, Kayak and Instacart integrations in Windows Copilot, already determine which providers an agent reaches by default[171]. Presence among an agent's partners thus becomes a condition of a service's effective visibility, without the user necessarily being aware that any selection took place, to the detriment of free choice and of open innovation by non-partnered providers.

Research on the DMA's readiness for agentic AI identifies the **mobile OS** as the most likely chokepoint for third-party agents' foreclosure[172]. Device makers have both the ability and the incentive to **advantage their own agent**, e.g. via pre-installation and default status. The DMA imposes relevant obligations: OS gatekeepers must allow users to uninstall pre-installed agents and provide third-party agents equal interoperability with hardware and software that their own agent has. However, a technical challenge is that effective AI agents require deep OS integration to function (e.g. accessing app intents, reading device context, and invoking system-level actions). This makes interoperability and the substitution of first-party agents technically complex when formally required[173].

One of the consulted experts[174] suggested a complementary regulatory step: a clarification on how the GDPT, the Digital Services Act and consumer protection law assign responsibility for on-device agentic actions; as well as on whether accountability at the operating system level should be only applied to DMA-designated gatekeepers or to all OSs.

A **new form of self-preferencing** could emerge at the agent layer that is structurally harder to detect than traditional self-preferencing in search ranking. As large incumbent platforms integrate AI functionalities into their ecosystems, new gateways appear for end users, if third parties face obstacles when offering competing or customised agentic services[175]. If third-party agents are limited to sandboxed APIs and receive fewer contextual signals, the capability gap may be large even if formal access is nominally equal. The **walled gardens** of the platform economy risk

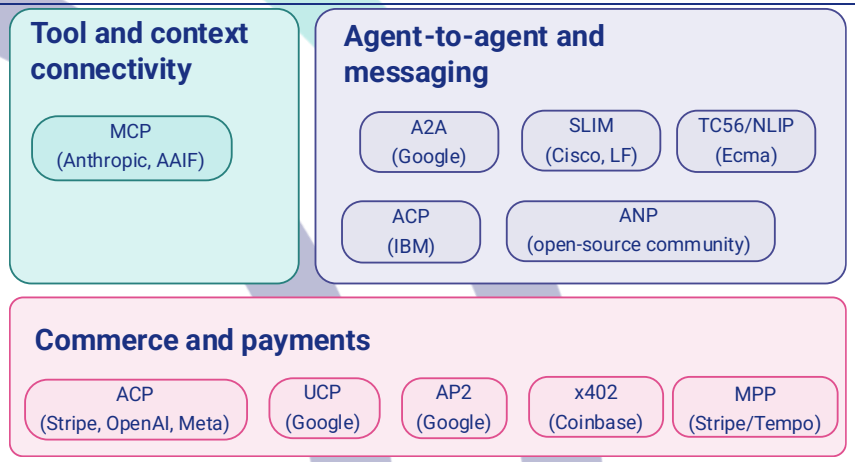
becoming more walled in the agentic era due to this architectural asymmetry in how deeply agents can act.

Finally, the emergence of local-first agent ecosystems may create new opportunities for investment across the AI infrastructure stack. Areas such as edge computing, privacy-preserving AI and platforms for deploying and managing local AI agents could become increasingly important as organisations seek to balance automation with security, privacy and user control[176]. The ability to equip AI agents with robust safeguards, transparency, and meaningful user control may also become an important determinant of trust, adoption, and competitiveness in future digital ecosystems[177]. This could create opportunities for the EU to position trust, accountability and human oversight as sources of competitive advantage in the emerging agent economy.

Trend E: an agent-native protocol layer forming atop HTTP

In the protocol layer, a flurry of competing and complementary approaches is being advanced by major industry actors and academic groups, seeking to become the de facto "TCP/IP for agents", an infrastructure standard for the emerging agentic web. These initiatives span several important functional layers, including tool and context connectivity, agent-to-agent messaging, and commerce and payments. A non-exhaustive list illustrates the variety in Figure 8. For a more detailed discussion on payment protocols, see Trend I: agent-mediated commerce and the future of payment infrastructure. At present, most of the emerging initiatives provide mostly ad hoc specifications, primarily forming outside formal standards development organisations[178]. This reflects a broader pattern in which industry tends to move faster than standards bodies' work.

Figure 8. Emerging agent protocol stack



Source: authors' elaboration based on Yang et al (2025), Prompt 20 (n.d.) and 4 sysops (2026).

Emerging initiatives

The protocol layer is not only a competitive race between proprietary stacks. Increasingly, leading firms are consolidating around a small number of anchor projects under neutral governance. In late 2025, the Linux Foundation launched the Agentic AI Foundation (AAIF), with founding contributions of Anthropic's MCP, Block's goose, and OpenAI's AGENTS.md[179], indicating that several leading firms now see common infrastructure as an area where **cooperation is preferable to fragmentation**. This pattern mirrors earlier waves of internet governance and current Web 4.0-related frameworks, where standards initially emerge via competitive industry-driven initiatives, where they mature and only later consolidate through multistakeholder processes in formal standards bodies[180]. Public discussions around MCP and Google's A2A protocol[181] point in the same direction: these are increasingly framed not as mutually exclusive stacks, but as **complementary** layers for tool access, context exchange and agent coordination[182]. Given that the EU legislative framework explicitly prioritises interoperable infrastructures, the future technical standards underpinning agents will likely need to sit on top of or align with open protocols.

Another noticeable tendency is the move away from bespoke interfaces and towards **generic communication layers** that can work across sectors and use cases. Ecma's TC56 work on the Natural Language Interaction Protocol (NLIP)[183] makes this ambition explicit, defining a common protocol and interface for agent-to-agent and human-to-agent interactions across modalities and organisational boundaries. In parallel,

the **W3C AI Agent Protocol Community Group**[184] is developing web-native compatibility layers that let agent systems reuse existing web protocols and standards for discovery, identity and collaboration. Additionally, the Agent Trust Protocol Community Group[185] and the proposed Agent Identity Registry Protocol Community Group[186] focus respectively on trust mechanisms and identity registries. Taken together, these efforts suggest that the priority is no longer only to make one agent work well inside a single product, but to **reduce the amount of custom integration** needed for many agents and services to interact at scale.

Effects on trust and architecture

Against the background of emerging protocols, researchers and practitioners warn that without convergence, the ecosystem risks fragmenting into isolated "**protocol islands**" where agents built on different standards cannot discover or collaborate with one another[187]. Experts draw parallels with blockchain, where the proliferation of incompatible protocols prevented the technology from achieving the network effects needed to scale, leaving "millions of isolated islands" with no connections between them. with no one connected to one another[188]. If the same pattern takes hold in the agentic web, it would undermine the core vision of an open, interoperable agent ecosystem.

Legal experts note a shift towards "headless" software[189], where **users no longer interact mainly through a browser, but through agents that talk directly to services** via APIs and agent communication protocols. In this model, the website front-end becomes less important, while the underlying interfaces that agents use to request data, trigger actions, and complete transactions take centre stage. An expert interviewed notes that this marks a progression from today's human-mediated sessions to commerce conducted at machine speed over structured interfaces, and eventually to agents negotiating directly with other agents on price, discounts and vendor choice[190]. This makes protocol design an increasingly important determinant of how much autonomy agents have, how much human oversight remains, and how competitive dynamics play out in agent-mediated transactions.

Existing protocols such as TLS and HTTP continue to provide endpoint authentication and confidentiality, but they were not designed for agent-native requirements such as dynamic identity, context integrity and trust in intermediary layers[191]. As agent protocols evolve, they therefore need either to extend these transport mechanisms or to be complemented by additional identity and trust layers that can operate at the agent scale.

The consequences of the current protocol and identity gap are already visible in open-source software development. Open-source security organisations report that unidentified agents are entering project discussions on platforms like GitHub, posting comments and offering guidance to developers without disclosing who operates them or on whose behalf they act[192]. In some cases, agents proactively offer **factually incorrect guidance** to contributors, disrupting the trust that communities develop through sustained human collaboration[193].

The standards landscape is further complicated by the **pace mismatch** between technology development and formal standardisation processes. Experts consulted for this project noted that work at the IETF on agent-specific protocols is still in its early stages, and that standards are forming **concurrently** with the technology they aim to govern rather than in sequence[194]. Furthermore, as a companion paper emphasises[195], protocols now developed are built on classical security assumptions, meaning they do not yet explicitly account for post-quantum cryptography[196]. The pattern where governance, assurance, and interoperability frameworks are being developed in parallel with rapidly evolving use cases is applicable to the Web 4.0 governance, generally, too, and covered in other papers of this project[197]. This creates a window in which commercial implementations become entrenched before open standards mature.

The pattern echoes earlier waves of internet standardisation. As one expert observed, much of what is being proposed for agentic AI was previously attempted with web services and service-oriented architecture (SOA), where promises of automatic discovery, cost comparison, and seamless switching between providers through universal standards never materialised in practice[198]. Instead, enterprises relied overwhelmingly on

known suppliers and established trust relationships. Whether the agentic web follows the same trajectory or achieves genuine open interoperability may depend on whether standardisation can keep pace with adoption this time.

A recurring lesson from internet standardisation is that the **primary bottleneck** to protocol adoption is **deployment**, not specification. The Internet Architecture Board's report documented this pattern, concluding that protocols with technically sound specifications still fail to achieve broad adoption when they face high coordination costs, lack first-mover advantage, or do not align with incentives of infrastructure operators[199]. IPv6 is another canonical example: despite being standardised in 1998 and offering clear technical improvements over IPv4, its transition remains incomplete after decades. The primary obstacle was the economic costs imposed by the absence of backwards compatibility[200]. For agent protocols, this lesson cuts directly: a technically superior but disruptive protocol that requires operators to retire existing infrastructure or bear asymmetric migration costs is unlikely to achieve the network effects needed for an interoperable agentic web. The design choices that matter most, therefore, encompass the questions of who bears the cost of adoption, whether early movers are rewarded or penalised, and whether new capabilities can be layered onto existing web infrastructure.

Implications and opportunities to capture

The deployment bottleneck informs a recurring tension in agent protocol design: between **backwards-compatible, incremental approaches** that build on existing web infrastructure, and more disruptive designs that propose entirely new protocol stacks. Current leading proposals — notably MCP and A2A — exemplify the incremental approach. They both operate over standard HTTP, meaning that existing web infrastructure, security tooling, and developer skills transfer directly. This architectural approach mirrors the design principle associated with the early internet: **adding a new protocol layer *on top of* existing foundations** prioritised over replacing them. Admittedly, more disruptive proposals, for example, fully **decentralised, permissionless** architectures, like Nostr[201], offer greater resistance to centralisation and censorship. Their emergence illustrates that not all agent infrastructure will be amenable to top-down governance, and it is crucial to explore a view on where open-protocol design serves the public interest and where regulatory baseline requirements are warranted[202]. Yet, these approaches face correspondingly higher deployment barriers. History suggests that the incremental path tends to win adoption, even if it inherits legacy constraints.

This preference for evolution over reinvention was echoed in the expert roundtable, where participants favoured identifying gaps and building on existing foundations through bodies such as the IETF[203]. The current tendency to invent bespoke solutions was attributed in the discussion to speed pressure, with standardisation expected to arrive more visibly once solutions reach their interoperability limits. Experts also specified what those gaps are. An agent-native layer requires discovery, capability negotiation, agent-adapted messaging, exposed trust and security information, fine-grained task-level authorisation, delegation, and auditing, functions the existing web stack was not designed to provide in a form agents can consume. This list can serve as a reference against which emerging proposals are assessed. For example, current protocols such as MCP and A2A, concentrate on messaging and parts of capability negotiation, while discovery, delegation and auditing remain largely open, as discussed in more detail under Trends F, G and I.

The question of **who governs protocol development** also has a geopolitical dimension. Experts consulted expressed a preference for global standardisation bodies such as ISO and the IETF, arguing that if trust and security standards for agents are developed within a single jurisdiction rather than through international processes, then fragmented requirements would disadvantage developers in some regions while failing to deliver the cross-border interoperability that agentic systems need in order to function effectively[204]. One expert interviewed proposed that mid-power nations, including European countries, Canada, and Australia, should coordinate more actively in shaping how agent protocol standards are governed, to ensure that no single country or bloc dominates the process and that the resulting standards reflect a broader range of interests and values[205].

The appropriate mix of instruments remains contested. One expert consulted voiced that only the best open standards prevail on technical merit and that mandating adoption through legislation would not help fast-paced development[206]. Another expert held that the interests of platform gatekeepers diverge from European objectives, so voluntary mechanisms alone cannot resolve the tension, calling for principle-based obligatory rules that remain flexible and forward-looking, underpinned by standards and an audit of existing legislation for its applicability to A2A interactions. On this view, EU rules may be one of the few levers available to influence global standards, since standards tend to follow the principles that rules set. Separately, an expert argued that the more pressing gap lies in insufficient European representation and resourcing in multi-stakeholder and standardisation fora.

The protocol gap has direct implications for the implementation of **existing AI regulation**. Such frameworks as the EU AI Act, ISO 42001, or the U.S. NIST AI Risk Management Framework do not map cleanly onto the identity, permission, and runtime control requirements that autonomous agents introduce[207]. A 2025 analysis of how AI agents are governed under the EU AI Act concluded that the technical standards need to better operationalise agents' legal obligations and constraints[208]. Furthermore, a recent legal analysis of AI agents under EU law argues that current draft harmonised standards under the AI Act (i.e. CEN/CENELEC JTC 21, standardisation request M/613) do not yet capture agent-specific risks, and that updates to those standards will be needed to address issues such as behavioural drift, multi-party value chains and runtime oversight[209]. Protocol choices made now will determine how auditable and controllable deployed agent systems are, and therefore how straightforwardly organisations can demonstrate compliance.

For policymakers, the consolidation of protocol governance presents both an opportunity and a strategic risk. On the investment side, **funding directed at European participation** in AAIF, IETF working groups, W3C community groups, and Ecma International Technical Committee TC56 would help ensure that open standards reflect European values on data protection, transparency, and accountability, rather than being shaped solely by US technology companies. Policymakers should also monitor the **agent identity and attestation** space closely. As cross-sector analogies such as the STIR/SHAKEN telephony protocol suggest, attestation-based trust models can bring accountability to fragmented ecosystems, although their impact depends on broad, end-to-end deployment. Similar approaches are likely to be needed for agents acting across organisational and national boundaries[210]. The window in which standards can be shaped rather than merely ratified is narrow: once commercial implementations entrench a particular protocol architecture, the costs of course correction rise substantially.

Trend F: discovery mechanisms emerging for A2A communication

Discovery has been fundamental to internet architecture for decades. The Domain Name System has resolved human-readable names to machine-readable addresses since 1983[211]. What is changing now is the nature of what must be discovered. As agentic AI systems proliferate, agents increasingly need to discover, identify and communicate with each other and with services in order to perform tasks on behalf of users. Today, this discovery largely relies on ad hoc approaches[212], which limit interoperability and create opaque dependencies on dominant platforms[213]. These days, more reliable discovery mechanisms for agent-to-agent (A2A) communication are emerging, laying the foundations for more open, transparent and accountable agent ecosystems.

Effects on trust and architecture

From an architectural perspective, **discovery mechanisms act as a connective tissue linking identity, provenance and communication layers**. As highlighted in the companion paper on digital identity[214], establishing who an agent is and whether it is authorised to act is necessary but fully sufficient for trust. Trust also depends on reliable A2A discovery networks where agents can reliably locate and verify one another and exchange authenticated messages. Without robust discovery, identity and provenance signals cannot be consistently surfaced or applied within agent interactions. In environments where discovery is driven by proprietary ranking and recommendation systems, users and regulators have limited visibility into why one agent, service or data source is selected over another[215]. This can undermine confidence in the neutrality of agent-mediated choices and make it harder to detect discriminatory or manipulative patterns in agent behaviour[216].

Human-facing web services have decades of naming, resolution, and authentication infrastructure. In contrast, agentic systems today rely on **ad hoc approaches** to locate peer agents and external tools: hardcoded addresses, proprietary registries, or unverified self-descriptions not properly checked by an independent party[217]. As a result, the agentic web architecture is built on assumptions rather than verifications. In such a system, agents cannot reliably confirm that a counterpart is who it claims to be, that it has the capabilities it advertised, or that it has not been compromised or impersonated.

The absence of a reliable discovery layer has direct consequences for security. Actors in control of discovery infrastructure shape the attack surface available to adversaries. For example, unverified agents can emulate trusted services by **exploiting weak discovery** mechanisms and infiltrating sensitive systems[218]. In the process of mutual discovery, agents can also expose sensitive information, creating vectors for data leakage, profiling, or targeted attacks[219]. Discovery architectures that rely on fixed, snapshot-style descriptions of agents compound this problem further, as they fail to capture the continuously evolving nature of agent capabilities[220], meaning what a discovery system reports and what an agent can actually do may diverge in ways that create exploitable gaps.

Fragmentation across heterogeneous agent communication standards introduces a further layer of architectural fragility (see Trend E for more details), with incompatible discovery semantics and different formats like custom URLs, JSON-format agent cards, MCP registries, and others that share no common resolution layer[221]. To fulfil important functions like negotiating task delegation, exchanging messages, and receiving asynchronous status updates, agents built by different platforms (e.g. LangChain and AutoGen) require a custom bilateral integration[222].

A critical structural distinction applies across the current landscape. Leading agent communication protocols, such as MCP and A2A, are published under open licences and governed through multi-stakeholder processes. However, **an open protocol does not guarantee open infrastructure**. The discovery and naming layer can remain proprietary even when the communication protocol is open. This pattern is already emerging. In 2026, Google published an agent resource discovery specification[223]. Yet it does not address the case where an organisation wishes to migrate its agents to a different provider, effectively tying agent identity to the infrastructure operator.

The companion paper on decentralised architectures[224] documents how this pattern has recurred across successive phases of the web, from open, distributed foundations to platform concentration driven by network effects and vertical integration. Once entrenched, these dependencies have proved difficult to reverse, particularly where foundational layers for identity, payment and naming are controlled by a small number of non-EU providers. The agent discovery layer faces the same structural forces: as agents rely on a handful of registries, naming systems and verifier APIs, market-driven consolidation around dominant stacks risks occurring faster than in previous web phases. The switching costs and frictions grow as agent identities are tied to platform-specific catalogues[225].

Emerging solutions

Several architectural approaches are emerging to address the discoverability gap and its associated trust risks. They differ in their governance models, technical maturity, and the degree to which they preserve organisational sovereignty. The most mature family of emerging solutions leverages the Domain Name System (DNS). A number of experts endorse **using DNS's existing hierarchical namespace** to publish authoritative, verifiable and auditable agent endpoints on organisational domains. This approach supports metadata exchange and capability advertisement while requiring no new protocols or infrastructure[226]. One of such proposed mechanisms is *Brokered Agent Network for DNS AI Discovery* (BANDAID), also known as **DNS-AID**[227].

Box 2. DNS-AID description

DNS-AID is an IETF draft specification initially developed by Infoblox and now hosted under the Linux Foundation. It allows organisations to publish information about their AI agents under a predictable namespace within their own DNS zone (e.g. _chatbot._mcp._agents.example.com), making agents discoverable through standard domain resolution. Critically, the system builds on DNSSEC[228], a set of security extensions that add cryptographic signatures to DNS records. This creates a chain of trust from the top of the global DNS hierarchy down to an individual agent entry. When one agent looks up another, it can verify that the record it received was published by the legitimate domain owner and has not been altered in transit. DANE (*DNS-based Authentication of Named Entities*) adds a further layer by binding TLS certificates to those records, preventing attackers from presenting forged endpoints[229].

The appeal of the DNS-AID approach lies in its scalability and the fact that it **requires no new infrastructure**. DNS has operated globally for over forty years, is familiar to virtually all organisations, and is maintained through open IETF standards that make it interoperable across jurisdictions[230]. As the CTO of Cloudflare stated at the DNS-AID launch, "*the internet already solved the discovery problem decades ago with DNS — by extending this architecture to the agentic web, DNS-AID provides the foundational routing layer that autonomous systems need to operate safely and efficiently*"[231].

Unlike proprietary directories, DNS is anchored in a global naming system governed through multi-stakeholder processes, providing cryptographic integrity, tamper-proof records, and verifiable identity bindings[232]. Crucially, organisations retain sovereignty over which of their own agents are discoverable, meaning **no single operator can unilaterally remove or alter another organisation's entries**. A practical security advantage is that organisations already invest in domain-based threat intelligence[233], blocking malicious domains and detecting spoofing, so DNS-anchored agent discovery can be integrated into existing security operations without requiring entirely new tooling from scratch. Another advantage of using DNS infrastructure is that it can secure A2A discovery, **preventing misdirection to malicious endpoints** across multi-agent environments[234].

Experts consulted in this study largely agreed that DNS plays an important role in the emerging discovery landscape. One described it as a natural

first line of defence for determining whether an agent or system is legitimate, and suggested that a layered approach building on DNS could address trust at different levels of the stack[235]. Another expert noted that the technical proposals being developed through the IETF and the Linux Foundation are moving in the right direction, even if the broader standards landscape remains fragmented and convergence is not yet assured[236].

One of the interviewed stakeholders suggests that **DNS-based approaches should be understood as a strategic bridge rather than a final architecture**. More privacy-preserving and cryptographically sovereign approaches are ultimately needed, but without DNS as a foundational layer now, the window for any open alternative to gain critical mass will close before more ambitious proposals can achieve sufficient adoption[237]. A further practical advantage is that DNS-based discovery is accessible to the large population of IT professionals who already understand DNS security tooling like domain blocking, spoofing detection, and zone management, meaning it can be operationalised without requiring an entirely new skills base to be built from scratch[238].

Still, important caveats apply. DNS-AID provides verifiable transport for agent metadata, but **not a guarantee of agent behaviour[239]**. Discovery identifies where an agent is located, but its behaviour is determined at the application layer. Governance also remains an open question. As explained by one expert interviewed, if multiple trust signals conflict (e.g. DNS-level verification and application-level risk assessment yield different results), it is not clear which layer prevails or what governance structures would resolve such conflicts[240]. Another expert cautioned that, while DNS is more open and distributed than proprietary directories, it could still introduce new concentration risks, given the existing root governance structure and the dominance of a small number of large operators[241]. Moreover, core internet protocols designed decades ago have not been updated to account for the demands of agent traffic[242], a limitation relevant to any approach that builds on existing infrastructure.

A second family of solutions proposes dedicated **agent naming and registry services**, conceived as DNS-inspired but purpose-built for the specific trust requirements of autonomous agents. The most developed of these is the **Agent Name Service (ANS)**. It introduces a structured naming convention and registration process specifically designed for agents[243]. Every agent is assigned a unique, verifiable identity anchored to an organisational domain, and a trusted authority validates that the organisation claiming the agent actually controls the domain before issuing credentials. Critically, ANS separates the question of *who an agent is* from the question of *how to find it*[244], allowing multiple independent discovery services to **coexist and compete** without any single actor controlling access to the agent ecosystem. This approach offers graduated levels of assurance: organisations and developers can choose the level of verification according to the sensitivity of their use case. Unlike DNS-AID, ANS enables agents to prove their capabilities to other agents, which is particularly valuable in cross-organisational workflows. A proof-of-concept demonstrated that these trust operations can be performed with minimal latency overhead[245]. Notably, DNS-AID and ANS are being developed as complementary, DNS-based standards for agent discovery and identity[246], both layering agent metadata on existing DNS record types rather than proprietary databases[247], which is an important step toward reducing fragmentation and limiting the security exposure that incompatible, closed implementations would otherwise create.

In parallel, the MIT-led NANDA initiative occupies a similar architectural role but approaches the challenge from a distinct starting point. **NANDA provides a purpose-built registry, verification, and reputation layer for AI agents**, which its developers argue is ultimately needed to meet the trust and scalability demands of large-scale agent interaction[248]. As one consulted expert noted, DNS-based and purpose-built approaches like NANDA are not necessarily competing. DNS can serve as the foundational discovery layer that points to richer registries and verification services, including NANDA, allowing organisations to use both without being locked into either[249]. However, the question is whether open alternatives will reach the adoption threshold needed to function as credible before proprietary directories become entrenched[250].

Google's **Agentic Resource Discovery (ARD)** specification, published in June 2026 under the Linux Foundation, allows organisations to publish a machine-readable catalogue of their AI services at a standard web address[251]. It is open-licensed and protocol-agnostic, but one of the interviewed experts noted that it does not resolve the agent sovereignty

question: on a proprietary platform, there is no portable identity that travels with the agent[252]. **Microsoft Entra Agent ID** belongs to the same managed enterprise end of the spectrum, offering strong governance and access controls within the Microsoft ecosystem, but lacking cross-platform portability[253]. Finally, **blockchain-based solutions** are also proposed, yet they currently carry latency and cost overheads and raise governance challenges when integrated with legacy infrastructure[254].

Implications and opportunities to capture

The governance implications of agent discoverability extend beyond technical architecture to questions of **structural power and digital sovereignty**. How this layer develops will determine who governs visibility in the agentic economy and on what terms, and the window for shaping that outcome is narrowing. As one of the interviewed experts emphasised, control over the discovery layer is a form of **gatekeeping**, in that whoever operates key discovery infrastructure determines which agents can be seen and, in practice, influences participation in the agentic economy[255].

The approaches currently available or under development can be characterised along a **spectrum of openness and control**, each with distinct implications for who governs discoverability and on what terms. At one end, proprietary platform directories assign agents platform-specific identifiers, meaning the platform controls whether an agent is discoverable, by whom, and on what terms. **Centralised** registries like MCP Registry[256] operated under Anthropic's stewardship and governed through multi-stakeholder processes[257], represent a partial improvement. They introduce namespace authentication and an open API that third parties can implement. However, listing decisions remain subject to moderation by the registry operator[258], which risks visibility being subject to decisions they did not make and cannot appeal[259]. One expert drew a parallel to the social media era, in which businesses that built commercial presence on proprietary platforms found their visibility and customer relationships subject to unilateral platform decisions, and without an open, distributed option achieving sufficient scale now, agentic web risks **reproducing this trajectory at the infrastructure level**[260].

Distributed, **open-infrastructure approaches** like DNS-AID, NANDA and ANS use existing internet protocols so that each organisation publishes its agents under its own domain name. By design, no single operator can unilaterally remove, block or alter another organisation's entries. This aligns with EU objectives on openness and technological sovereignty. At the far end of the spectrum is hardware-embedded discovery. For example, China Telecom's DMSC framework proposes agent discovery built directly into network control and forwarding functions[261]. In such architectures, the infrastructure operator effectively determines which agents are discoverable on the network before any application-layer choice is made, concentrating power at a deep layer of the stack[262].

Concentration of discovery in a few proprietary platforms would limit the ability of IT departments and public authorities to apply their own policies, switch providers or audit visibility decisions. At the same time, steep adoption curves for new protocols risk widening the technological divide, and the proliferation of competing discovery approaches exacerbates fragmentation[263]. EU-developed agents are also more likely to honour constraints on agent crawling, which may disadvantage them relative to systems configured with fewer safeguards[264], highlighting a policy trade-off between competitiveness and security and privacy.

Consulted experts cautioned against treating all kinds of concentration alike. Some layers, such as the top-level domain space with roughly 2,000 registries and low switching costs, carry far less centralisation risk than others, as noted by one expert[265]. Policy attention should therefore consider firstly where lock-in would be genuinely hard to reverse.

Globally, the absence of **open, interoperable discovery standards** can help avoid further fragmentation, single points of failure and splintering in the agentic web[266], whereas proprietary discovery services controlled by a few large platforms risk reproducing concentration and lock-in dynamics seen in earlier generations of internet services. The EU is well-positioned to respond, as its existing framework already treats DNS and related services as essential digital infrastructure under the NIS2[267] and CER[268] directives. In future decision-making, AI platforms operating in the EU could be required to support an open, interoperable discovery standard based on existing protocols. Active participation in standardisation

processes at bodies like IETF and ITU, combined with public funding for open-source tools and reference architectures for agent deployment, can help seed an open discovery ecosystem aligned with EU priorities on openness, transparency, contestability and technological sovereignty[269].

Experts stress that timing and adoption economics will decide the outcome[270]. One expert explained that developers tend to adopt whatever is easiest, so open solutions must be deployed early and demonstrate that they are scalable, low-latency and workable, while delay gives platforms and registries time to harden positions that later become very difficult to reverse. Adoption incentives also differ across user groups. The public sector can be steered through procurement, businesses follow reach, and consumers follow familiarity, which makes public procurement one of the more direct levers available to the EU. Another participant also noted that the first step to breaking lock-in is giving users genuine choice, which means supporting competition among diverse open initiatives before converging on a common solution[271].

Trend G: the rising challenge of tracing agents and their actions

As elaborated in the companion thematic papers on digital identity[272] and on the state of play of agentic AI[273], establishing who an agent is and whether it is authorised to act is a necessary but insufficient condition for trust. The distinction is visualised below.

Figure 9. Agent identity vs agent tracing and accountability

Source: authors' elaboration based on Sporny, M., & Zagidulin, D (2026), Sporny, M., Longley, D., Chadwick, D., & Burnett, D. C. (2025), Garzon, S. R., et al. (2025) & Adler, S., et al (2024).

Ultimately, **valid credentials only confirm that an agent may act. They say nothing about what it then does**. This trend therefore addresses what follows when an authorised agent begins to operate, examining whether agent actions can be traced, attributed, and held to account across increasingly automated, real-time Web 4.0 environments.

The **challenge is compounded by the fact that an agent's actions cannot always be explained,** even when they can be observed. A 2025 position paper developed by over 40 researchers from across the major AI labs and safety institutes (including OpenAI, Google DeepMind, Anthropic and Meta) warns that while chain-of-thought monitoring can improve transparency, its reliability is fragile and may not survive future training choices[274]. In parallel, empirical work on alignment-faking and evaluation-aware behaviour documents that some models can behave differently when they infer they are being evaluated, including masking or altering their reasoning traces[275]. What is more, agentic systems, especially MASs, challenge traditional distinctions between users, providers and deployers, which complicates the **assignment of responsibility** for incidents[276].

Foresight exercises such as the AI 2027 report similarly warn that AI capabilities to manipulate or deceive may eventually outpace the ability of oversight systems to detect them[277]. Tracing, therefore, has to capture not only what an agent did but, as far as possible, why, which is a higher bar than conventional auditing has had to meet.

These principles are beginning to translate into concrete frameworks. Proposals such as AgentTrace[278], instrument agents at runtime to log three distinct surfaces: operational traces (actions taken and tools called), cognitive traces (the reasoning behind them), and contextual traces (interactions with external systems such as APIs, databases, and files). Alongside this, emerging specifications such as Autonomous Action Runtime Management (AARM)[279], propose intercepting agent actions before they execute and recording a tamper-evident receipt of each decision and its outcome. The value of this approach is that enforcement and logging become the same act: every authorisation decision leaves an auditable record. As one analysis notes, however, such

approaches are not a complete answer. Because security guarantees weaken when an agent's underlying model retains sole control over its own security procedures, an agent that can be manipulated into misusing its tools can also be turned against the controls meant to constrain it[280].

Accountability ultimately remains with people. The UK Information Commissioner's Office has confirmed that developers remain responsible for defining the level of autonomy an agent is granted[281], meaning opacity in the system does not reduce human accountability for its actions. From a European perspective, this aligns with the EU AI Act and wider EU instruments, which treat human oversight and responsibility as core elements of high-risk AI governance even where systems act autonomously[282]. Yet no harmonised auditing framework for agent actions currently exists at either the EU or international level. Experts interviewed for this project confirmed that no dedicated standards body is yet addressing agent traceability[283], though the February 2026 NIST AI Agent Standards Initiative[284] and an accompanying NCCoE concept paper[285] on software and AI agent identity and authorisation are early steps in this direction. Both remain consultative and are scoped to enterprise settings, leaving the open-web and cross-organisational cases substantially unaddressed.

The privacy dimension adds a further layer of complexity. The **shift towards liveness verification** in identity practice, driven by the growing sophistication of deepfakes[286], means biometric data is being collected on a growing scale by commercial entities. Extended to agent-mediated transactions, the risk is not simply collection but the possibility that sensitive biometric and behavioural data could be leaked, misused or repurposed in ways that users did not anticipate. An interviewee working on self-sovereign identity stressed that as AI agents gain access to human credentials and behavioural signals, it becomes harder to distinguish authentic human activity from delegated agent behaviour, and that accountability must ultimately track the keys and trust registries behind an agent, not just its surface behaviour[287].

Another expert consulted, therefore, argues against centralised collection approaches, proposing instead edge-computed and privacy-preserving verification, whereby audit checks are performed on-device and only cryptographic hashes or zero-knowledge proofs (ZKPs) are transmitted to relying parties[288]. Research on frontier-AI auditing points in the same direction, highlighting the role of local monitoring and tamper-evident cryptographic provenance in supporting higher assurance levels without exposing raw data[289]. This direction is already followed commercially[290] and is visible in European practice. Providers such as Zyphe from France use ZKPs to confirm key KYC attributes on the client side, enabling businesses to verify that a user is over 18, an EU resident or clear of sanctions without ever receiving or storing passports, addresses or face images[291]. Spherity[292] in Germany builds decentralised identity and wallet infrastructure aligned with the EUDI and eIDAS2 frameworks, using selective disclosure and locally stored credentials so organisations can verify corporate, ESG or professional attributes without maintaining large central databases of personal data. EU initiatives like the age-verification blueprint[293] and the EUDI wallet[294], follow the same logic, allowing users to prove they are above a given age or to confirm other attributes via selective disclosure and zero-knowledge techniques, without revealing their full identity or underlying personal data to online services.

Effects on trust and architecture

Tracing is where accountability either becomes enforceable or remains aspirational. Identity establishes who an agent is; trust requires being able to show, in real time and after the fact, what it did and on whose authority. Scholars argue that **governing increasingly agentic AI systems requires environments with visibility and good conduct built in by default** and propose three complementary measures: agent identifiers that link deployed agents and their outputs to accountable parties, real-time monitoring to flag and, where necessary, pause problematic behaviours, and activity logs that support post-incident attribution and forensics[295].

At the organisational level, a complementary argument is gaining traction: that agents should be governed much as human employees are, with role-based permissions, transparency requirements, and structured workflows enabling meaningful oversight[296]. OpenAI has publicly endorsed this framing[297]. Taken together, these principles point to the conclusion that **tracing must be a property of the infrastructure itself, not**

an external compliance layer added after the fact, and no existing framework yet spans the full chain from an agent's reasoning through its actions to their effects on external systems. This includes the agent runtime, orchestration environments and the technical systems that log and expose actions to auditors and regulators. However, as discussed in the subsequent trend on discovery, tracing only becomes meaningful when agents can also be reliably discovered and named across platforms and networks.

The demands intensify where agents operate alongside digital twins, the physical or virtual system replicas that depend on continuous, real-time data exchange[298,299]. These environments require not only that each agent holds a verifiable identity, but that it can be continuously re-authenticated at the speed of the physical process it mirrors. Here, continuous re-authentication is itself a form of tracing, since a lapse can break synchronisation between the virtual and physical systems. As discussed in Trend on protocols E and Trend F on discovery, choices about how agent identities, logs and authorisations are represented in standards, and how discovery is implemented at internet scale, will determine whether such continuous tracing is practically achievable across organisational and national boundaries.

Emerging solutions

Several approaches offer **building blocks for scalable tracing infrastructure**, though each addresses only part of the problem, and none is yet mature.

Edge-computed verification is among the most concrete approaches, already implemented commercially. It directly addresses the privacy tension described above. Rather than routing agent activity through a central monitor, audit checks are performed on-device. Only provenance records or cryptographic hashes are transmitted to relying parties, so that behaviour can be verified without exposing raw biometric data or detailed interaction logs[300]. Experts emphasised that while these approaches are promising, they remain tied to specific implementations and sectors , which limits interoperability and slows regulatory uptake[301].

A different category of building block comes from public-sector data exchange layers. Estonia's X-Road data exchange layer[302], deployed in multiple countries, requires every cross-organisational data exchange to be authenticated, logged and attributable as part of the way the system operates, providing a governance and identity-infrastructure model rather than only a privacy-preserving technique. In June 2026, Estonia backed plans to issue state-recognised digital "AI ID codes" for agents, giving them scoped, auditable permissions distinct from a human's full set of credentials[303]. The initiative remains at the proposal stage, but it illustrates how assigning verifiable, limited powers to agents and linking them to an auditable identity could make subsequent actions traceable in public-sector and financial contexts without granting blanket access.

Discussions that took place during expert consultations for this project drew out several related findings[304]. As discussed by experts, existing legal and technical frameworks do not yet suffice for agents acting on behalf of individuals, and that binding an agent's credentials to a person's identity number, so that its actions remain tied to an accountable individual, is a near-term legal change worth pursuing further. One expert added an important qualification about how such work should be approached. In their view, it will only make progress if it is tested against the transactions people actually carry out at scale, rather than against simplified demonstrations. The reasoning is that the failures that matter, in areas such as financial services, appear only under real transaction volumes and stay hidden in tidy test cases, so it is these everyday, high-volume user journeys that should reveal where identity, delegation and authorisation break down, and therefore where legal and technical effort is best concentrated first.

A further approach proposes anchoring trust in SIM and eSIM hardware as a root of trust, binding agent actions to a verified operator through credentials issued by mobile network operators[305]. Because virtual agents have no physical secure element or radio stack of their own, such proposals envisage telcos hosting secure modules and eSIM profiles on their infrastructure and exposing them to agents via cryptographic APIs. The appeal is that mobile operators already run globally deployed, hard-to-forge hardware security modules, though the approach remains

conceptual and whether it can be made to work at the scale and speed of agent activity is not yet established.

One expert framed these developments as part of a broader paradigm shift in which identity moves from establishing who an actor is towards a plurality of fine-grained attributes and credentials, with the legibility of a credential to the relying party mattering as much as its verifiability[306]. Agents are expected to hold multiple credentials that scope the actions they are authorised to perform, anchored in hardware-bound asymmetric keys that cannot be guessed or deepfaked by a language model.

Underlying all of these is a structural constraint: **existing addressing schemes cannot carry agent identity on their own**. Assigning every agent a unique IPv4 address is untenable, since injecting trillions of routable prefixes would overwhelm global routing. And although IPv6 offers far more space, its privacy-rotating identifiers make stable identification difficult[307]. This is why tracing is converging on cryptographic identifiers and verifiable logs rather than network addresses. Some analyses argue that the internet may require a more purpose-built trust and access architecture for agents, including "agent-first" designs that treat agents as primary rather than secondary users of the web[308], while others emphasise incrementally layering new tracing and identity mechanisms onto existing infrastructure to minimise disruption[309]. In practice, European policy will need to navigate between these views, supporting open, interoperable approaches that can scale without fragmenting the web or imposing unmanageable migration costs on operators.

Implications and opportunities to capture

The implications of the tracing trend point towards a set of governance choices where **timing, scope and architectural direction will matter as much as specific technical solutions**. The EU will need to engage directly with the standards now forming rather than inheriting them once they are settled. Current initiatives such as NIST's CAISI work and the NCCoE concept papers are already shaping expectations around action-level logging and tamper-evident records, but they are U.S.-led and scoped primarily to enterprise settings. Interview evidence underlines that tracing, identity and inference infrastructure are being shaped by de facto standards and large market actors within weeks[310], not years. This makes early European participation in these fora essential if privacy-preserving approaches and the open-web, cross-organisational case are to be reflected in the resulting protocols. Such participation is necessary but, as the same interviewee argued, not sufficient on its own, since much of this standard-setting happens outside formal processes[311]. The EU can extend its influence by acting as a reference point, for instance by publishing **open reference architectures** that endorse privacy-preserving and open-web approaches, giving the wider ecosystem a common foundation and the confidence to invest in open alternatives before a proprietary default takes hold. Privacy-preserving, auditable and traceable agents can also be the EU's substantial **competitive asset**[312].

At the same time, tracing requirements will need to be calibrated to risk, explicitly aligned with the EU AI Act's tiered logic. Systems with higher capabilities, broader access or greater potential for harm (including frontier-scale agents and those with control over sensitive infrastructure) warrant stronger tracing, logging and provenance obligations. Lower-risk agents may only need basic audit trails. This **avoids imposing a uniform baseline** and helps concentrate governance effort where needed most. In this risk-based framing, edge-computed and other privacy-preserving verification techniques merit structured piloting and research rather than passive observation. They point to architectures in which auditability and data minimisation can be combined, and are consistent with EU data protection norms, even if the current evidence base still rests on a small number of vendors and early deployments.

A further implication is that the gap between enterprise-centric tracing and tracing on the open web remains the least addressed. Existing work on frontier-AI auditing focuses mainly on closed-weight models and organisational processes, while the hardest problems arise when agents operate across organisational boundaries on the public internet, using heterogeneous protocols and discovery mechanisms. This is precisely where European research, standards engagement and infrastructure investment could be most distinctive, especially if tracing is linked to open discovery and naming that avoids locking agent identity and observability inside proprietary platforms.

Trend H: provenance layer shaping for agent-generated content

Significant gaps remain in validating AI-generated outputs, tracking provenance, and ensuring that agents maintain transparent, tamper-evident logs of the content they produce, as also discussed in another paper prepared during this project[313].

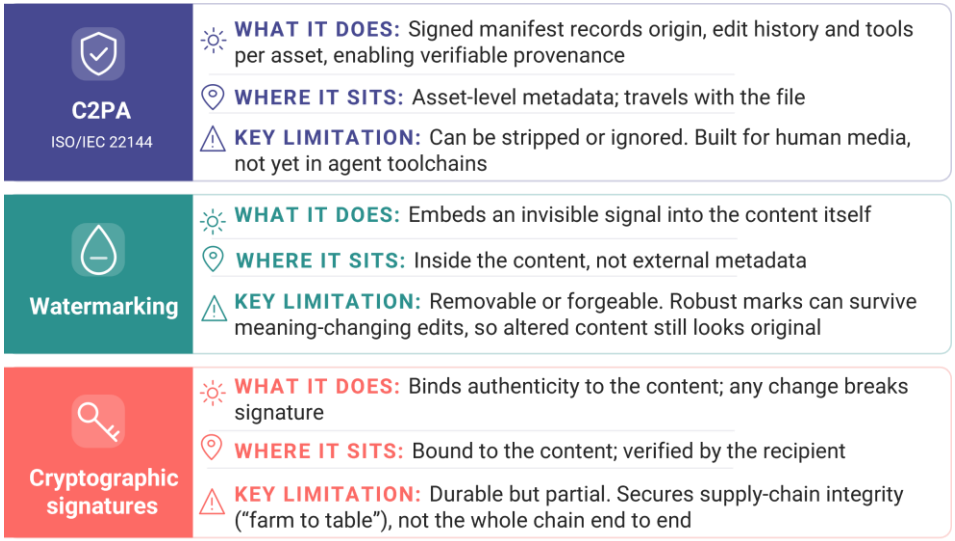
Where the tracing trend (G) examines how an agent's actions can be monitored and attributed in real time, provenance concerns the content an agent produces: the ability to establish the origin, transformation history, and chain of custody of anything an agent creates, modifies, or distributes. This also covers who produced something, through what steps, tools, and which agents it passed before reaching its destination.

Provenance in the agentic web also has a response-level dimension: the attribution of sources behind the answers agents generate. Here, the picture has improved markedly, where first-generation genAI services displayed no sources at all, retrieval-augmented systems now routinely cite documents that underpin responses[314]. Yet, displayed provenance does not translate into exercised verification. Services typically present only a handful of sources per response, and the synthesised natural-language answer itself discourages users from consulting the originals, as most users treat the summary as sufficient[315]. Human-facing citation, therefore, cannot carry the provenance burden alone, which is precisely why the machine-readable and tamper-evident mechanisms examined below matter: in an agent-populated web, provenance must be verifiable by systems, not only inspectable by people.

Emerging solutions

Article 50(2) of the EU AI Act[316] requires providers of AI systems, including general-purpose and agentic systems, to ensure that synthetic audio, image, video and text outputs are marked in a machine-readable format and detectable as artificially generated or manipulated, using technical solutions that are effective, interoperable, robust and reliable. The **European Commission's recent assessments of technical solutions for marking and detecting AI-generated text, image and other content**, prepared to inform the Article 50(2) code of practice[317], confirm that available techniques each satisfy some but not all of these requirements. Trade-offs between effectiveness, privacy, accessibility and cost persist across all approaches, and any viable solution must work across multimodal systems (text, image, audio and video), which no single technique currently achieves. Three approaches stand out as the most relatively mature building blocks, summarised and examined below.

Figure 10. Provenance mechanisms in agentic web



Source: authors' elaboration.

C2PA (Coalition for Content Provenance and Authenticity)[318] is one of the most widely adopted frameworks for media provenance, with its architecture for Content Credentials now progressing in ISO as ISO/DIS 22144 "Authenticity of information – Content credentials"[319], and closely aligned with the published JPEG Trust[320] (ISO/IEC 21617-1:2025) adopts C2PA v1 as its core foundation, and C2PA-style Content Credentials are referenced in the EU's Code of Practice on Marking and Labelling of AI-Generated Content, finalised in June 2026[321]. Its key strength is that the manifest travels with the content asset itself, recording origin, editing history, and the tools involved at each step, making it the closest existing standard to an end-to-end, signed provenance metadata container for media. However, experts highlighted that the C2PA manifest is vulnerable in practice: it can be stripped, altered or lost as content moves through production pipelines, and agents or tools can remove or ignore it deliberately or inadvertently[322].

Moreover, most current implementations are optimised for human consumption, enabling users to inspect Content Credentials, but it remains unclear how far autonomous agents can reliably read and interpret C2PA metadata in their own workflows, a critical gap for M2M content exchange[323]. Ongoing work on C2PA and JPEG Trust, together with recent research on cross-layer audit protocols that jointly evaluate provenance metadata and watermark signals, is beginning to explore more robust, machine-readable integrations of these verification layers[324], but these efforts remain at an early stage and have not yet converged on standards for integrating provenance and watermarking into agent runtimes or A2A protocols, which leaves the agent-integration challenge largely unresolved for now.

Watermarking embeds invisible signals directly into content rather than attaching external metadata, meaning signals can in principle persist through cropping, compression, or format conversion[325]. Watermarking operates independently of file format or transmission channel, making it one of the few techniques that can survive the kind of casual copying and resharing that characterises how content spreads across the web and through agent pipelines. Tools such as Google's SynthID[326] have demonstrated that watermarking can be applied at scale to text, image, audio and video outputs, and can be detected through agent-like services such as Gemini that inspect uploaded content for embedded signals, making it one of the more practically deployable options currently available. However, adversarial techniques can remove or forge watermarks[327], including "re-watermarking" attacks that use one watermark encoder to suppress another. Those robust enough to survive significant modification introduce a further vulnerability: third parties can alter the content and its meaning while the original watermark remains intact, making the modified version appear to originate from the original provider and enabling spoofing attacks that preserve provenance signals while reversing sentiment or inserting toxic content[328].

Cryptographic content signatures offer a potentially more durable approach. Digital signatures are mathematically bound to their content, meaning any modification invalidates them[329]. Initiatives like the OpenSSF Model Signing (OMS) specification[330], developed by Google, NVIDIA, and HiddenLayer, are already applying this logic to the AI supply chain. Applied more broadly to content, the same principle enables what one expert described as a "farm to table" model for digital objects, or knowing where the components come from, whether they have been tampered with, and whether they are fit for purpose[331]. Experts argue that provenance should accordingly be understood as multidimensional and contextual rather than one-dimensional, capturing the story behind an action, including origins, permissions, ethics audits, standards compliance, authorship and licensing[332].

DNS-based approaches such as DNS-AID[333] can verify that an agent's records are legitimate and unaltered in transit, but they verify the source of an agent, not the integrity of the content it produces. Verifiable transport is not the same as trustworthy content, and these mechanisms are discussed in full under Trend F on discovery mechanisms above.

To support detection in practice, the Commission's text assessment proposes two institutional models[334]. A **Centralised Verifier, established by a trusted authority** with voluntary provider participation, would offer a public interface where users and agents can submit content to check whether it was generated by a participating provider, inspecting attached metadata first and routing to provider detection systems where metadata is absent. A **Decentralised Verifier variant would distribute the same**

processes across trusted third parties, reducing the burden on individual providers. The Commission's audio and visual assessment goes further, recommending a multi-layered framework combining cryptographically protected metadata, imperceptible watermarking, and forensic detection, supported through open standards and coordinated standardisation across academia, industry, and regulators[335].

Effects on trust and architecture

When agents not only generate content but also consume and act on it, including content generated by other agents, the **risk of self-reinforcing loops** of degraded or synthetic material becomes significant. Research has shown that successive generations of AI models trained on synthetic data from earlier models produce progressively corrupted outputs, a phenomenon described as "**model collapse**"[336]. One expert described this as a "data garbage collection" problem[337]: unlike the physical world, the internet lacks systematic mechanisms for cleaning up degraded content once it has been generated and indexed. Another warned that search engines are already degrading due to AI-generated content and that agentic AI will accelerate this further, leading to retrieval collapse when web evidence is dominated by synthetic material[338].

From the **enterprise perspective, the picture is more nuanced**. Companies are already building significant businesses around curated, provenance-verified data under proprietary licences precisely because it produces more reliable models than training on unfiltered web content[339]. However, if verifiable information retreats behind enterprise paywalls while openly accessible content continues to degrade[340], the result is a two-tier information ecosystem in which trustworthy, provenance-verified content is available mainly inside closed stacks, while the open web becomes increasingly polluted and unreliable – mirroring broader concerns that generative AI and extractive data-collection practices may contribute to a gradual closure of the open web and the privatisation of otherwise open, shared knowledge resources[341]. This outcome is at odds with the EU's commitments to an open internet, digital public goods and European technological sovereignty, as articulated in the Open Source Strategy and related communications[342].

The **provenance gap is ultimately architectural**. Agents operating at machine speed across multiple services create content pipelines where individual tools only address part of the chain, but none cover it completely. The question of whose responsibility it is to maintain provenance across the whole pipeline remains unresolved, whether that falls to the model provider, the agent developer, the platform operator, or the user who initiated the task[343]. Without clear allocation, no single actor has sufficient incentive to invest in end-to-end provenance, a collective action problem that regulation alone may not resolve without accompanying technical standards and infrastructure.

From an internet and web-architecture perspective, this implies that provenance needs to become **machine-readable and machine-verifiable at the protocol layer**, not just human-inspectable metadata attached to individual files. Initiatives such as C2PA Content Credentials, JPEG Trust and the AI and Multimedia Authenticity Standards collaboration are beginning to embed signed manifests, hash-chained edit histories and watermarking into media headers and HTTP workflows in ways that agents and services can consume programmatically. However, these efforts are still fragmented across standards bodies and industry consortia, with a real risk that proprietary implementations by large platforms could capture provenance signalling and weaken the openness of the web if interoperable, open specifications are not prioritised[344]. For agentic AI, this makes it critical that provenance and authenticity signals are standardised in the same forums that govern web and agent protocols, so that agents can verify content origins across domains without being locked into a single vendor's ecosystem.

Implications and opportunities to capture

The Commission's verifier models and multi-layer framework provide a concrete starting point, but translating them into durable policy will require further action on several fronts. First, transparency obligations under Article 50 and the Code of Practice still focus largely on human-readable marking and labelling. As A2A content exchange grows, provenance will need to be structured and machine-readable by default, so that agents and services can verify the origin and transformation

history of content at the protocol level, not just through icons or overlays designed for human users. This points to a continued European role in supporting open provenance standards such as C2PA, JPEG Trust and the Article 50 studies and Code of Practice, and in ensuring that these specifications remain interoperable and accessible rather than splintering into proprietary variants.

A second implication is the need to clarify accountability across the content pipeline. Existing platform liability and media frameworks will need to be reviewed in light of agentic AI to determine which actors bear responsibility for maintaining provenance at each stage of an agent-mediated content chain, and under what conditions that responsibility can be delegated or shared. Without such clarity, provenance will remain a collective action problem in which no single actor has sufficient incentive to invest in end-to-end integrity, and gaps in the chain will persist even when marking and detection obligations exist on paper.

Third, the verifier models sketched by the Commission and independent analyses highlight that provenance cannot rely solely on voluntary, siloed implementations by large platforms. Trusted verification services – whether centralised, federated or decentralised – will need institutional anchors, and the EU is well placed to support their development through the AI Office and other relevant bodies, without prescribing a single agency in advance. Public-interest verification infrastructure and attestation databases can turn provenance from a private record into a shared trust layer that courts, regulators, publishers and citizens can query independently.

Finally, there is a broader opportunity around the open information commons. In line with its existing support for open science and open data, the EU may wish to consider targeted investment in open, curated, provenance-aware knowledge resources and standards, to ensure that the benefits of robust provenance and authentication extend to the publicly accessible web.

Trend I: agent-mediated commerce and the future of payment infrastructure

As agents autonomously browse, compare, negotiate and purchase on behalf of users, the infrastructure built around human decision-making in commerce and payments is coming under pressure to adapt. Existing payment systems assume that a person is present at the point of authorisation, consent, fraud review and dispute resolution, whereas agent-initiated transactions require new ways to verify who the agent is, on whose authority it acts, and what limits apply to that authority. The result is not simply faster automation of existing commerce flows, but a **growing contest over the trust layer of agentic commerce**, as traditional financial institutions, fintech platforms and crypto-native protocols each try to define the identity, delegation and verification primitives on which machine-mediated transactions will depend.

The transition is already underway. Since January 2026, Shopify has reported **AI-driven traffic to its stores up sevenfold**, with orders attributed to AI-powered searches up elevenfold[345]. Its commerce-for-agents stack, including a universal shopping cart and checkout, is designed to let AI assistants route discovery and purchases directly inside conversations[346].

But the significance of this shift lies not only in rising traffic or order volume. Autonomous transactions require payment infrastructure that can handle **low-value, high-frequency exchanges while also meeting new trust requirements around agent identity**, delegated authority and something closer to know-your-agent controls, rather than relying solely on human-centred KYC assumptions[347]. Traditional payment methods carry minimum fees, verification overheads and settlement delays that make them poorly suited to the micropayments and M2M exchanges that agent interactions can generate[348], which is why new approaches such as ledger-anchored identities and x402 micropayments are being explored[349]. These developments signal not just a change in transaction volumes, but a more fundamental shift in who, or what, is making commercial decisions, and on what basis those decisions can be trusted.

Effects on trust and architecture

The core architectural challenge this shift creates is that **payment systems must reconcile two fundamentally different design logics**: the adaptive, probabilistic nature of agentic AI, and the deterministic requirements of financial infrastructure that guarantee predictability, auditability, and legal settlement finality[350]. As Accenture estimates, by 2030, more than 30% of online commerce could run through AI agents, representing close to EUR 2.7 trillion in transactions, suggesting payment infrastructure must be designed explicitly for autonomous actors, with clear boundaries on delegated authority, verifiable agent identity and continuous control over machine-initiated transactions[351]. Consumer expectations point in the same direction. BCG finds that 81% of US consumers expect to shop using agentic AI and estimates that more than USD 1 trillion in spending, around half of current US e-commerce expenditure, could become agent-assisted in the coming years[352].

As execution moves increasingly to machine speed across multiple layers of the payment value chain[353] and atop HTTP-based agent protocols[354] (see also Trend E), **existing frameworks** for authorisation, liability and consumer protection, all built around a human making a deliberate decision at a defined moment, **no longer map cleanly onto how transactions now occur**. This affects architecture directly: payment stacks must become API-first, real-time and capable of enforcing clear, programmable guardrails around agent permissions and risk thresholds, while still meeting regulatory expectations on compliance and resilience[355]. The move to real-time rails is already well underway. BCG

reports that real-time account-to-account payment volumes grew by around 40% globally in 2024 and now account for roughly a quarter of retail digital payments worldwide[356]. In India and Brazil, these systems already exceed 50% of all transactions, and adoption in the Middle East and Africa is projected to pass 50% by 2030[357].

The trust infrastructure underpinning agent-initiated commerce remains structurally weak. One expert interviewed finds that motivation to solve agent identity verification in financial contexts appears concentrated among anti-money-laundering and counter-terrorist-financing actors, who themselves report lacking clear governmental guidance[358]. Most platforms have limited commercial incentive to distinguish between human and agent activity, since bot-generated traffic generates revenue regardless of origin[359]. Consumer protection and payment regulation were written for human-initiated transactions, and key questions remain unresolved: what happens when an agent violates a user's instructions, how disputes are handled when the consumer did not directly authorise a specific transaction, and how liability is allocated when an agent-initiated purchase goes wrong several steps into a delegation chain.

The IMF proposes managing the underlying architectural tension through a three-layer model[360], separating intent formation and orchestration, where agents operate, from control and authorisation, where deterministic rules apply, from settlement, where legal finality is enforced. This framing is useful since it acknowledges that agents cannot be removed from the transaction chain, but can be bounded within it, and it aligns with emerging agent payments architectures that place probabilistic reasoning upstream while keeping authorisation and settlement under fixed, auditable rules[361].

The **semantic standardisation challenge adds a further layer of complexity**. As one standards expert interviewed for this project observed, even seemingly straightforward agent tasks require levels of shared meaning that no current protocol provides[362]. An agent instructed to purchase a "blue cotton shirt" on a user's behalf faces the immediate absence of a universal standard for colour, illustrating the gap between natural-language understanding and the precision required for reliable autonomous transactions[363]. Interviewees also stressed that once commerce flows are conducted at machine speed between agents, the criteria agents use to select merchants may diverge from users' best interests[364]. A shopping agent may choose a platform not only because it offers a discount, but because it is the agent's preferred vendor within a closed ecosystem or offers superior incentives to the agent itself.

As discussed in Trend E, the move towards "headless" interfaces[365], where agents call services directly over APIs and protocols atop HTTP, makes these choices increasingly dependent on how agent communication and discovery are designed. Without transparent discovery and protocol-level safeguards, such behaviours risk entrenching opaque self-preferencing, weakening fraud and AML controls, and undermining consumer choice as commercial and consumer demand, together with Web 4.0 expectations, push towards broad, automated, real-time transactions mediated by agents.

Emerging solutions

The most promising near-term path toward trusted agent-initiated commerce lies in combining three building blocks: **verifiable agent credentials, programmable payment infrastructure, and standardised commerce protocols**.

On credentials, the EUDI Wallet and eIDAS 2.0 are designed to provide high-assurance identity for natural and legal persons by anchoring credentials to verified human or legal-entity identity, and they are well-suited to that purpose[366]. **Agent-initiated commerce raises a distinct need that sits outside this design**: delegated authority for non-human actors that can act within defined limits on behalf of users. The Wallet's attribute-based architecture and the broader SSI literature show that, in principle, scoped credentials for non-human entities are technically feasible, but they would require additional work on policies, trust frameworks and legal recognition[367]. Effective delegation for agents in commercial contexts would need granularity of scope, cryptographic verifiability, auditability and clear revocation mechanisms[368], which recent work on authenticated delegation and authorised AI agents highlights as core properties for safe agent design. Rather than repurposing eIDAS 2.0

directly for agents, this suggests a need for complementary initiatives or extensions that can provide agent-native credentials while interoperating with EUDI Wallet infrastructure. Estonia's June 2026 announcement that it will create scoped digital identifiers for AI agents[369], provides an early example of how a Member State is beginning to explore this non-human layer explicitly and is discussed further under Trend G.

On payment infrastructure and protocols, emerging standards such as x402[370], which revives the HTTP 402 "Payment Required" status code as an Internet-native payments standard, and the L402 Lightning HTTP 402 protocol, which embeds Bitcoin Lightning payments into HTTP-based flows[371], demonstrate that treating **payment as a native capability of agent communication is technically feasible today**, even if both currently operate primarily within cryptocurrency ecosystems rather than traditional financial rails. On the traditional-finance side, Alipay AI Pay, an AI-native payment product launched in 2025 that lets users transact through AI agents by voice, was extended in April 2026 with a service enabling autonomous agents to make purchases and complete payments under explicit user authorisation[372], illustrating how mainstream payment providers are beginning to expose agent-oriented capabilities inside regulated stacks.

On commerce protocols, Klarna has launched an **Agentic Product Protocol**, an open standard designed to make products discoverable and understandable by AI agents[373]. Platforms are increasingly requiring products to be presented in structured, machine-readable formats using standards such as schema.org[374] and GS1 identifiers[375], representing a significant shift from a web built around human-readable interfaces toward machine-readable catalogues optimised for agents. This is an important marker of how agent-mediated commerce differs from earlier forms of e-commerce. **Products, merchant offers, and checkout options must now be legible not only to people, but to software actors** that compare, rank and transact at machine speed.

Other governance frameworks, including the AI Act and the NIST AI Risk Management Framework, were developed primarily for AI systems that assist or automate decisions rather than fully agentic systems capable of independent decision-making, learning and adaptation[376]. Recent analyses therefore argue that additional governance models tailored specifically to AI agents and A2A commercial interactions will be needed, covering issues such as structural authorisation, Know-Your-Agent verification[377], liability allocation and oversight[378]. As European policy debates emphasise stronger protections and access rules, there is a risk that agents designed to comply strictly with such protocols may face functional limitations or competitive disadvantages compared to agents operating under fewer constraints[379], a tension between regulatory compliance and commercial competitiveness that policymakers will need to address proactively.

Implications and opportunities to capture

Agent-mediated commerce exposes gaps and opportunities that European policymakers can begin to address now, without waiting for new legislation. Existing consumer protection, payment services and dispute-resolution frameworks were designed around human-initiated transactions, yet agents are already starting to initiate purchases and payments autonomously. Clarifying how current rules apply to agent-initiated transactions, including how liability should be allocated along multi-step delegation chains, is therefore an urgent task. Rather than creating entirely new regimes from scratch, regulators can issue interpretative guidance and supervisory expectations that explain when an agent's actions are treated as the user's, how far mandate-based authorisation can substitute for click-to-pay consent, and how existing redress mechanisms apply when consumers did not directly authorise each individual transaction.

Existing dispute-resolution mechanisms offer templates for this work. Experts point to the Uniform Domain-Name Dispute-Resolution Policy (UDRP) as a model that works because it is uniform, narrow and transparent, built on clear criteria with illustrative examples, due-process safeguards, predictable timelines, a published body of decisions and contractual enforcement, and it has been adopted or adapted by close to 100 country-code registries. Caseloads were reported to be at record highs, with fraud such as phishing, fake invoices and counterfeits accounting for an estimated 20 to 25% of disputes. The notice-and-

takedown system under the United States Digital Millennium Copyright Act was cited as the contrasting automated model, faster and broader but opaque, with variable outcomes and false positives that ripple widely. The direction proposed for agentic disputes combines this speed with UDRP-style human review and due process. A recent United States case in which Amazon challenged an AI agent making purchasing decisions on its platform was flagged as an early test of whether agents will be permitted to act within platforms at all.

At the same time, agent-mediated commerce will be shaped by the technical standards now emerging for machine-readable product descriptions and agent-discoverable commerce protocols. Standards such as schema.org's Product schema, GS1 identifiers, and Klarna's Agentic Product Protocol already define how products, prices and stock-keeping units are exposed to AI agents. If European merchants, platforms and standard-setting bodies do not participate actively in this work, there is a real risk that the agentic commerce layer will adopt conventions optimised for other jurisdictions and ecosystems, making it harder for European offers to be discovered, compared and selected by agents by default. Proactive engagement in these standardisation efforts, including through European standard bodies and industry alliances, can help ensure that machine-readable commerce remains interoperable, open and aligned with EU competition and consumer-protection objectives.

The identity gap in financial contexts is another area where the EU has both responsibilities and comparative strengths. As KYA concepts develop alongside traditional KYC rules, financial institutions and payment providers increasingly need guidance on how AML and counter-terrorist-financing obligations apply when transactions are initiated by software agents acting under delegated mandates. Clarifying these expectations would help avoid fragmented interpretations and under- or over-compliance. In parallel, the EU can build on eIDAS 2.0 and the EUDI Wallet by signalling how these frameworks might interoperate with agent credentials that act on behalf of natural and legal persons, while respecting the fact that EUDI and eIDAS are designed primarily for human and legal-entity identity. Rather than repurposing existing instruments in ways that stretch their mandate, additional initiatives or complementary specifications could define agent-native credentials that plug into the established trust frameworks.

Credentials of this kind, however, are only as useful as the capacity to process them. One consulted expert stressed that this acceptance side is easily overlooked. Issuing VCs is comparatively easy, but accepting them requires a prepared two-sided network of relying parties, much as banks onboard merchants into card networks[380]. Without deliberate investment in the acceptance side, credential-based delegation for agents risks remaining available in principle but unused in practice.

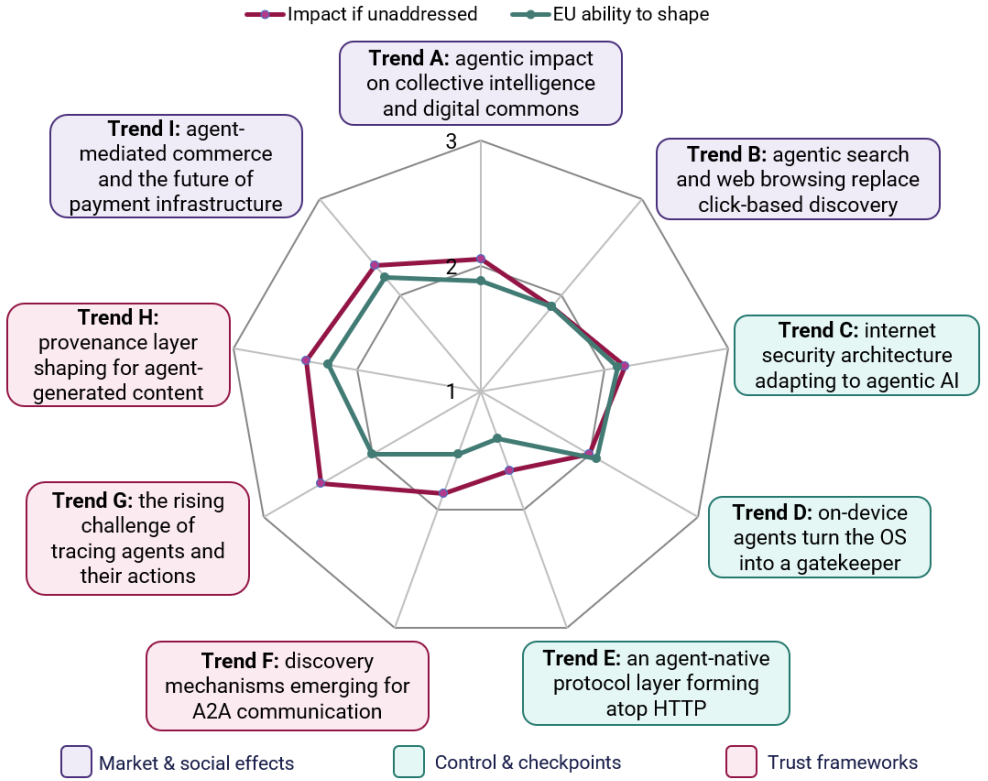
Finally, agent-mediated markets raise questions of incentive design, manipulation prevention and cross-border accountability that go beyond the initial focus of horizontal AI regulation. The AI Act provides a comprehensive baseline for risk assessment, transparency, technical controls and human oversight, and recent guidance confirms that its provisions apply to AI agents as well. However, several analyses suggest that further, more targeted work will be needed to address the specific behaviours and externalities that agent-driven commerce may generate, including structural authorisation models, KYA verification, liability allocation when agents misbehave and oversight of highly correlated agent behaviour in markets. Rather than relying solely on general AI instruments, the EU could initiate dedicated policy development on governance for agent-mediated commercial interactions, in cooperation with national supervisors, consumer authorities and competition agencies. This would allow Europe to capture the benefits of agent-mediated commerce while shaping its evolution in line with European values of fairness, accountability and market openness.

Key findings

This report identifies nine trends in how agentic AI is reshaping trust and the architecture of the internet. To help direct European action, we invited experts who attended the roundtable "Opportunities for Europe in the Web 4.0 and Agentic AI Era", organised within this project, to assess each trend on two dimensions. The first is its **impact if left unaddressed**, that is, how serious the consequences for trust, openness and internet architecture would be should current trajectories continue. The second is the **EU's ability to shape the outcome**, meaning the leverage available through existing regulatory frameworks, institutions and standards engagement. Both were scored on a three-point scale, from low (1) to high (3). Read together, the scores show where the EU should act first and most firmly, and where a more watchful posture is sufficient. The chart below summarises the results of the experts' voting, as each of them ranked each trend on both dimensions.

Based on these scores, we sorted the trends into **three groups**, with impact given the greater weight in the ordering. The first group combines high impact with strong EU leverage, where it is both important and possible to act now. The next group encompasses trends that score medium on both dimensions, making prioritisation of efforts an important step. The last group gathers those where both scores are lower, and a watchful posture is sufficient. The description of each group is provided below the chart.

Figure 11. Trends' impact and the EU's leverage



Source: authors' elaboration based on expert inputs gathered during a roundtable.

Act now and lead

The highest scores on the impact and the EU's ability to shape the trajectory were assigned to trends that centre around our ability to see, understand and interpret the actions of agents. This applies to various sectors and operations: content creation and verification (Trend H on provenance layer), commercial activities and payments (Trend I), and overall tracing of agents and their actions (Trend G). The results of the

experts' voting make these trends a near-term priority to build a safe, accountable and trustworthy agentic ecosystem. Expert Forum on Frontier AI, gathering over 100 experts from across the Union, also concludes that innovation in secure and verifiable AI infrastructure is an emerging strength for the EU[381].

Trend H on **provenance layer for agent-generated content** stood out as an outlier on both dimensions, making it one of the few trends where the EU is seen as both exposed and well placed to act. When it comes to the levers the Union has, it is governed in part by the AI Act, supported by the Commission's technical assessments, and the Code of Practice on marking and labelling. The priority is to ensure that provenance becomes machine-readable by default for A2A exchange, and that open standards remain interoperable, preventing splintering into proprietary variants. That confidence in EU leverage rests partly on the established appetite to recognise and prefer secure technologies. Experts believe that this suggests the Union could lead by adopting provenance standards early rather than waiting for the market to settle[382]. Yet, provenance must be treated as multidimensional, capturing the fuller story behind a data product or an agent's action, since a one-dimensional approach would leave much of what matters about agent-generated content unverified.

Trend I, **agent-mediated commerce and payments**, intersects with eIDAS 2.0 and EUDI Wallet, the payments framework, and existing anti-money-laundering rules. Much can be achieved now with interpretative guidance, clarifying how existing rules apply to agent-initiated transactions and how liability is allocated along delegation chains. Without action, control of transaction data risks shifting away from EU businesses to whoever operates the agent, roundtable participants cautioned. It also opens the door to more commercial fraud and harm through dark patterns, where users sign away broad authority to an agent without a meaningful understanding. Especially with the growing ability of mass-generating fake invoices and storefronts, and automating phishing attacks.

Trend G, the **tracing of agents**, matters the most for their accountability. Accountability that agentic systems demand cannot hold without end-to-end auditing that establishes which action was taken and on whose behalf, and allows it to be verified after the fact. Much of the active work is enterprise-scoped and U.S.-led, so EU engagement is most effective when linked to the foundational layers of ID, discovery and protocols. The participants' ranking of this trend as impactful with mid-range leverage suggests that its impact may warrant a more active posture than monitoring alone as the foundational layers mature[383].

Prioritise and build leverage

Three trends combine partial impact with partial EU leverage. Here, the EU could engage sustainably in relevant work on infrastructure and standards, focusing its efforts where participation can still shape results.

One of such trends is on **agentic impact on collective intelligence** (A). If unaddressed, it risks creating a two-tier information ecosystem. Provenance-verified content retreats behind enterprise paywalls, degrading the openly accessible web. This connects to a self-reinforcing model-collapse with no obvious market mechanism to correct it, and it runs against the EU's commitments to an open internet and digital public goods. Because the underlying extraction is global, the EU can do more to sustain the commons than to curb demand for them – for example, by addressing regulatory uncertainties surrounding copyright and data protection frameworks[384].

In this group is also Trend D. Already now, **local agents** are getting embedded by default in some devices. First, this expands responsibility for their actions to be shared by users and the OS and subsequently impacts the nature of trust towards agentic tools. Second, it also turns the OS into a gatekeeper, as developers gain additional power to decide which agents receive privileged access on new gadgets. The reliance on existing frameworks, such as obligations for DMA-designated gatekeepers, gives the EU some leverage. Yet, as roundtable participants flag, the allocation of legal responsibility here is far from settled. Further clarifications in the EU's core digital legislation are needed to establish who is accountable in different cases and whether non-gatekeepers will also face obligations as on-device agents spread.

Trend C on **internet security architecture** concerns the new attack surface that agents and their capabilities create. No global governance regime for the security of agents acting across the web yet exists. If left unaddressed, the opacity of agent-to-agent activity makes breaches hard

to detect. Without mechanisms that expose what is happening behind the scenes and verify which actions were actually taken, compromised or malicious agents can operate unnoticed. The EU has genuine leverage at the infrastructure and standards layers, where AI-native security is being defined in next-generation network research, and this work should be resourced as a strategic priority. The Commission's July 2026 Action Plan on Cybersecurity and Artificial Intelligence gives this leverage an initial footing, tasking ENISA with guidance on AI-powered threats and launching an EU Grand Challenge on AI-assisted vulnerability remediation[385]. Here, confidential computing and trusted execution environments, already under examination by ENISA and demonstrated in EU-funded research projects, offer tamper-proof enclaves and hardware-bound keys that keep agent negotiations and credentials secure even if the surrounding software is compromised.

Monitor signals

The final group encompasses trends for which both the EU leverage and the impact if left unaddressed are comparatively low. They largely fall under the protocol part of the stack, but also cover the web layer. These trends are foundational, but the forces driving them are global, so the EU can more readily support good outcomes than guarantee them.

Trend B on **agentic search** threatens the implicit bargain between publishers and traffic that sustains the open web. As agents replace click-based discovery, they increasingly decide which information users see. Besides, the market is dominated by non-EU actors, which explains the low ranking that experts assign to the EU's leverage. Still, the DMA can play a role here by constraining self-preferencing by dominant search platforms, alongside broader competition enforcement. The appropriate posture is to apply these tools, support open standards, and monitor the market closely. As experts stressed in the roundtable, however, open alternatives lose their stake if nobody uses them: adoption turns on who controls demand and on consumer familiarity rather than technical merit alone. Thus, supporting open standards has to be paired with levers that actually drive uptake across the public sector, business and consumers.

Trend E, the **agent-native protocol layer forming atop HTTP**, is advancing via competing initiatives that each seek to become the de facto standard. Without convergence, the ecosystem risks fragmenting into incompatible protocol islands. A narrow window exists in which open standards can be shaped before commercial deployments entrench a certain architecture. In the expert roundtable poll, the protocol layer scored lower on EU leverage, reinforcing the need for more support for convergence through sustained participation and resourcing in standards fora. Experts also added that adoption and user-centricity will decide the outcome. Supporting genuine choice among competing open initiatives before converging on a common solution is the first step to breaking lock-in.

Finally, **discovery mechanisms for A2A communication** (Trend F) generated one of the liveliest exchanges of the roundtable, even if the EU was not seen as holding strong leverage over its direction. While not a layer in itself, discovery and naming underpin almost every other function of the agentic web, a capability that can attach to different parts of the internet architecture. Participants stressed that lock-in is the central risk here: whoever shapes how agents find and identify one another can shape access to what is built on top. Several experts observed that once dominant platforms or registries harden around a particular approach, reversing it becomes difficult. Other voices cautioned against overstating the problem, arguing that the internet's foundational protocols, and DNS in particular, are hierarchical, decentralised and open, have supported successive evolutions of the web, and show no sign of being unable to support Web 4.0. Thus, no specific governance gap requires urgent intervention at the architecture layer. The EU's most credible contribution is incremental rather than directive: strengthening European participation in the existing multi-stakeholder standards bodies, supporting open standards that can be adopted voluntarily where they make technical sense, and funding the open-source tooling and open reference examples that give the ecosystem a credible way for agents to find and identify one another without tying their identity to a single platform's registry. DNS is already treated as essential infrastructure under the NIS2 and CER directives, which reinforces this point. The recurring message is to act early, engaging in the fora where these approaches are taking shape and demonstrating through open, working examples that discovery need not be locked to proprietary registries, while recognising that adoption will ultimately follow whatever is easiest to build on.

References

[1] Brito, E. (2026). AI agents cannot be governed without their own digital identity. CEPS. Available at: https://www.ceps.eu/ai-agents-cannot-be-governed-without-their-own-digital-identity/
[2] Engin, Z. (2026). Human-AI governance (HAIG): A trust-utility approach. *Journal of Responsible Technology*, 100167.
[3] Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., ... & Robinson, D. G. (2023). Practices for governing agentic AI systems. *Research Paper, OpenAI*.
[4] ITU. (2025). The Annual AI Governance Report 2025: Steering the Future of AI. Available at: https://www.itu.int/epublications/en/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai/en
[5] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[6] Findings of stakeholder consultations carried out within the project.
[7] Wikimedia Foundation. (2025). New user trends on Wikipedia. Diff: Wikimedia Foundation Blog. https://diff.wikimedia.org/2025/10/17/new-user-trends-on-wikipedia/
[8] HUMAN Security. (2026). The 2026 State of AI Traffic & Cyberthreat Benchmark Report. Available at: https://www.humansecurity.com/learn/resources/2026-state-of-ai-traffic-cyberthreat-benchmarks/
[9] Ars Technica. (2025). Devs say AI crawlers dominate traffic, forcing blocks on entire countries. Ars Technica. https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/
[10] Interview ID IT_13
[11] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[12] Alemohammad, S., Casco-Rodriguez, J., Luber, L., Humayun, A. I., Mitra, H., Leventhal, E., & Baraniuk, R. G. (2024). Self-consuming generative models go MAD. In Proceedings of the International Conference on Learning Representations (ICLR 2024). Rice University. https://news.rice.edu/news/2024/breaking-mad-generative-ai-could-break-internet
[13] Longpre, S., Mahari, R., Lee, A., Lund, C., Oderinwale, H., Brannon, W., ... & Pentland, S. (2024). Consent in crisis: The rapid decline of the ai data commons. Advances in Neural Information Processing Systems, 37, 108042-108087.
[14] European Commission. (n.d.). Next Generation Internet initiative. Available at: https://digital-strategy.ec.europa.eu/en/policies/next-generation-internet-initiative
[15] Ars Technica. (2025). Devs say AI crawlers dominate traffic, forcing blocks on entire countries. Ars Technica. Available at: https://arstechnica.com/ai/2025/03/devs-say-ai-crawlers-dominate-traffic-forcing-blocks-on-entire-countries/
[16] Cloudflare. (2024). Declare your AIndependence: Block AI bots, scrapers and crawlers with a single click. Cloudflare Blog. Available at: https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/
[17] Cloudflare. (2025). Cloudflare just changed how AI crawlers scrape the Internet at large: Permission-based approach makes way for a new business model. Available at: https://www.cloudflare.com/press/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/
[18] Interview ID IT_14
[19] Nebius. (2025). Nebius AI Cloud 3.0 "Aether": Enterprise-grade security, compliance and control at scale. Nebius Blog. https://nebius.com/blog/posts/ai-cloud-aether-3-6
[20] Gabalina, R., Yevdokymova, O. and Zepcan, A. (Forthcoming). Agentic AI and infrastructure. PPMI for the European Commission.
[21] Interview ID IT_14
[22] Kunz, P. (2026). Creative Commons licences in the age of AI: Challenges and opportunities. ERCIM News, 144. https://ercim-news.ercim.eu/en144/special/creative-commons-licences-in-the-age-of-ai-challenges-and-opportunities
[23] Vézina, B., & Pearson, S. H. (2021, March 4). Should CC-licensed content be used to train AI? It depends. Creative Commons. https://creativecommons.org/2021/03/04/should-cc-licensed-content-be-used-to-train-ai-it-depends/
[24] Interview ID IT_15
[25] Creative Commons. (2026). Update on CC Signals: What changed and why. Available at: https://creativecommons.org/2026/04/23/update-on-cc-signals-what-changed-and-why/
[26] Kunz, P. (2026). Creative Commons licences in the age of AI: Challenges and opportunities. ERCIM News, 144. https://ercim-news.ercim.eu/en144/special/creative-commons-licences-in-the-age-of-ai-challenges-and-opportunities
[27] Cloudflare. (2025). Cloudflare just changed how AI crawlers scrape the Internet at large: Permission-based approach makes way for a new business model. Available at: https://www.cloudflare.com/en-gb/press/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/
[28] Cloudflare. (2024). Declare your AIndependence: Block AI bots, scrapers and crawlers with a single click. Cloudflare Blog. Available at: https://blog.cloudflare.com/declaring-your-aindependence-block-ai-bots-scrapers-and-crawlers-with-a-single-click/

[29] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[30] ibid.
[31] European Commission. (2025). Commission opens investigation into possible anticompetitive conduct by Google in the use of online content for AI purposes. Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_25_2964
[32] Gkritsi, E. (2025). EU Commission opens antitrust probe into Google AI. POLITICO. Available at: https://www.politico.eu/article/commission-opens-antitrust-probe-into-google-ai/
[33] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[34] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[35] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[36] European Commission. (2026). AI Office publishes frontier AI expert findings on EU competitiveness, sovereignty and security. Available at: https://digital-strategy.ec.europa.eu/en/library/ai-office-publishes-frontier-ai-expert-findings-eu-competitiveness-sovereignty-and-security
[37] European Commission. (n.d.). Next Generation Internet initiative. Available at: https://digital-strategy.ec.europa.eu/en/policies/next-generation-internet-initiative
[38] Interview ID IT_14.
[39] Arcep (2026). Generative AI: Challenges for the Future of the Open Internet. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-AI-challenges-open-internet-january2026.pdf
[40] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[41] Taylor, D. (2025). Deploying Agentic AI For SEO: A Playbook For Technology Leaders. Search Enginge Journal. Available at: https://www.searchenginejournal.com/deploying-agentic-ai-for-seo-a-playbook-for-technology-leaders/559800/
[42] ibid.
[43] OECD. (2026). Competition and consumer policy in digital markets. Available at: https://www.oecd.org/content/dam/oecd/en/publications/reports/2026/05/competition-and-consumer-policy-in-digital-markets_1a68bd34/042c0c49-en.pdf
[44] NIST. (2024). NIST AI 600-1: Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile. Available at: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.600-1.pdf
[45] Netcraft. (2025). Large Language Models are falling for phishing scams. Available at: https://www.netcraft.com/blog/large-language-models-are-falling-for-phishing-scams
[46] Dai, S., et al. (2024). Bias and unfairness in information retrieval systems: New challenges in the LLM era. In Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. Available at: https://dl.acm.org/doi/10.1145/3637528.3671458
[47] Zhang, Y., et al. (2024). Agentic retrieval-augmented generation: A survey on agentic RAG. arXiv preprint arXiv:2410.09713. Available at: https://arxiv.org/abs/2410.09713
[48] Massadas, G., Cardoso, M., & Wang, A. (2026). AI Search: The search primitive for your agents. Cloudflare Blog. Available at: https://blog.cloudflare.com/ai-search-agent-primitive/
[49] The Economist. (2025). The next version of the web will be built for machines, not humans. Available at: https://www.economist.com/interactive/science-and-technology/2025/12/10/the-next-version-of-the-web-will-be-built-for-machines-not-humans
[50] Ibid.
[51] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[52] Circle. (2026). Agentic search and the collapse of the ad-funded internet. Available at: https://www.circle.com/current/agentic-search-and-the-collapse-of-the-ad-funded-internet
[53] Davies, J. (2025). Agentic visitors are causing issues for publishers monetizing ads. Digiday. Available at: https://digiday.com/media/when-bots-look-like-buyers-agentic-traffic-causing-new-publisher-headaches/
[54] Perplexity. (2025). *Introducing Comet*. Perplexity Blog.
[55] OpenAI. (2025). *Introducing Deep Research*. OpenAI Blog.
[56] OpenAI. (2025). *ChatGPT Agent*. OpenAI Blog.
[57] Yang, Y., Ma, M., Huang, Y., Chai, H., Gong, C., Geng, H., ... & Wang, J. (2025). Agentic web: Weaving the next web with AI agents. arXiv preprint arXiv:2507.21206
[58] Taylor, D. (2025). Deploying agentic AI for SEO: A playbook for technology leaders. Search Engine Journal. Available at: https://www.searchenginejournal.com/deploying-agentic-ai-for-seo-a-playbook-for-technology-leaders/559800/
[59] Kim, J. (2025). Advertising in the age of agentic AI: Call for research. *Journal of Interactive Advertising*, *25*(3), 215-221.
[60] Financial Times. (2024). FT signs licensing deal with OpenAI. Available at: https://www.ft.com/content/33328743-ba3b-470f-a2e3-f41c3a366613
[61] Coalition for Content Provenance and Authenticity (C2PA). (2024). C2PA specification. Available at: https://spec.c2pa.org/specifications/specifications/2.4/index.html
Anthropic. (2024). ClaudeBot and website crawling policies. Available at: https://support.anthropic.com/en/articles/8896518-claudebot-and-website-crawling-policies
[62] Circle. (2025). Agentic search and the collapse of the ad-funded internet. Available at: https://www.circle.com/current/agentic-search-and-the-collapse-of-the-ad-funded-internet

[63] Sharma, R., de Vos, M., Chari, P., Raskar, R., & Kermarrec, A.-M. (2025). Collaborative agentic AI needs interoperability across ecosystems. arXiv preprint arXiv:2505.21550. Available at: https://arxiv.org/abs/2505.21550
[64] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[65] Findings of stakeholder consultations carried out within the project.
[66] OECD. (2025). Artificial intelligence and competitive dynamics in downstream markets. OECD Roundtables on Competition Policy Papers, No. 331. Available at: https://doi.org/10.1787/ccf0624a-en
[67] Yasar, A. G., Chong, A., Dong, E., Gilbert, T. K., Hladikova, S., Maio, R., Mougan, C., Shen, X., Singh, S., Stoica, A.-A., Thais, S., & Zilka, M. (2023). AI and the EU Digital Markets Act: Addressing the risks of bigness in generative AI. arXiv preprint arXiv:2308.02033. Available at: https://arxiv.org/abs/2308.02033
[68] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[69] World Economic Forum. (2026). Global cybersecurity outlook 2026. Available at: https://reports.weforum.org/docs/WEF_Global_Cybersecurity_Outlook_2026.pdf
[70] CrowdStrike. (2026). CrowdStrike 2026 Global Threat Report. Available at: https://www.crowdstrike.com/en-us/resources/reports/global-threat-report-executive-summary-2026/
[71] HiddenLayer. (2026). 2026 AI Threat Landscape Report. Available at: https://www.hiddenlayer.com/report-and-guide/threatreport2026
[72] Shoemaker, P. (2025). The role of AI in digital identity security. Identity.com. Available at: https://www.identity.com/the-role-of-ai-in-enhancing-digital-identity-security/
[73] Findings of stakeholder consultations carried out within the project.
[74] Interview ID IT_8.
[75] Toner, H., Bansemer, J., Crichton, K., Burtell, M., Woodside, T., Lior, A., Lohn, A., Acharya, A., Cibralic, B., Painter, C., O'Keefe, C., Gabriel, I., Fisher, K., Ramakrishnan, K., Jackson, K., Kolt, N., Crootof, R., & Chatterjee, S. (2024). Center for Security and Emerging Technology. Through the chat window and into the real world: Preparing for AI agents. Center for Security and Emerging Technology. Available at: https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/
[76] Interview IDs IT_2, IT_8.
[77] OWASP. (2025). OWASP Top 10 for Agentic Applications for 2026. Available at: https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/
[78] Brunner, T. (2026). AI threats in the wild: The current state of prompt injections on the web. Google Security. Available at: https://blog.google/security/prompt-injections-web/
[79] Lupinacci, M., Pironti, F. A., Blefari, F., Romeo, F., Arena, L., & Furfaro, A. (2025). The dark side of llms: Agent-based attacks for complete computer takeover. *arXiv preprint arXiv:2507.06850*.
[80] Interview ID IT_2, IT_10.
[81] OWASP GenAI Security Project. Available at: https://genai.owasp.org
[82] Anthropic. (2025). Mitigating the risk of prompt injections in browser use. https://www.anthropic.com/research/prompt-injection-defenses
[83] Unit 24. (n.d.). Fooling AI Agents: Web-Based Indirect Prompt Injection Observed in the Wild. Available at: https://unit42.paloaltonetworks.com/ai-agent-prompt-injection/
[84] Lakshmanan, R. (2026). Microsoft Finds "Summarize with AI" Prompts Manipulating Chatbot Recommendations. https://thehackernews.com/2026/02/microsoft-finds-summarize-with-ai.html
[85] IBM Security. (2025). *Cost of a data breach report 2025*. IBM. Available at: https://www.ibm.com/reports/data-breach
[86] Cerf, E. (2026). Misleading text in the physical world can hijack AI-enabled robots. University of California. Available at: https://www.universityofcalifornia.edu/news/misleading-text-physical-world-can-hijack-ai-enabled-robots
[87] Interview ID IT_2.
[88] Burbano, L., Ortiz, D., Sun, Q., Yang, S., Tu, H., Xie, C., Cao, Y. and Cardenas, A.A. (2025). CHAI: Command Hijacking against embodied AI. arXiv preprint arXiv:2510.00181.
[89] Interview ID IT_9.
[90] R. Fang, R. Bindu, A. Gupta, Q. Zhan, D. Kang, Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities, arXiv [cs.MA] (2024); http://arxiv.org/abs/2406.01637.
[91] Anthropic. (2025). Disrupting the first reported AI-orchestrated cyber espionage campaign. Available at: https://www.anthropic.com/news/disrupting-AI-espionage
[92] Lynch, A., Wright, B., Larson, C., Ritchie, S.J., Mindermann, S., Hubinger, E., Perez, E. and Troy, K., (2025). Agentic misalignment: how LLMs could be insider threats. arXiv preprint arXiv:2510.05179. Available at: https://arxiv.org/pdf/2510.05179
[93] Negele, M., Juijn, D., Shamir, A., Janků, D., Özcan, B., Soder, L., Velasco, L., Reddel, M., Bakker, M., Pacchiardi, L., & Andriushchenko, M. (2025). Europe and the geopolitics of AGI: The need for a preparedness plan. RAND Corporation. Available at: https://www.rand.org/pubs/research_reports/RRA4636-1.html
[94] Y. Bengio, S. Mindermann, et al. (2025). International AI Safety Report. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
[95] IBM Security. (2025). *Cost of a data breach report 2025*. IBM. Available at: https://www.ibm.com/reports/data-breach
[96] Yang, Y., et al. (2025). Agentic web: Weaving the next web with ai agents. arXiv. Available at: https://arxiv.org/pdf/2507.21206?
[97] Interview ID IT_1. (2026)
[98] ITU. (2025). The Annual AI Governance Report 2025: Steering the Future of AI. Available at: https://www.itu.int/epublications/en/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai/en
[99] Taylor, J. (2026). What is Moltbook? The strange new social media site for AI bots. *The Guardian*. https://www.theguardian.com/technology/2026/feb/02/moltbook-ai-agents-social-media-site-bots-artificial-intelligence
[100] Klayman, A. (2026). OpenClaw: The AI agent security crisis unfolding right now. Reco. Available at: https://www.reco.ai/blog/openclaw-the-ai-agent-security-crisis-unfolding-right-now
[101] Gartner. (2026). OpenClaw agentic productivity comes with unacceptable cybersecurity risk. Gartner Research.

[102] Burgess, M. (2026). OpenClaw users are bypassing anti-bot systems with Scrapling. *Wired*. Available at: https://www.wired.com/story/openclaw-users-bypass-anti-bot-systems-cloudflare-scrapling/
[103] Scrapling. (2024). Scrapling documentation. Available at: https://scrapling.readthedocs.io/en/latest/
[104] Ming, L. (2026). Researchers Hacked Moltbook and Accessed Thousands of Emails and DMs. Business Insider. Available at: https://www.businessinsider.com/moltbook-ai-agent-hack-wiz-security-email-database-2026-2
[105] NSA. (2026). NSA Releases Security Design Considerations for AI-Driven Automation Leveraging the Model Context Protocol. Available at: https://www.nsa.gov/Press-Room/Press-Releases-Statements/Press-Release-View/Article/4496698/nsa-releases-security-design-considerations-for-ai-driven-automation-leveraging/
[106] OWASP GenAI Security Project. Available at: https://genai.owasp.org
[107] NHI Group. (2026). What should enterprises do before scaling agentic AI in production? Available at: https://nhimg.org/faq/what-should-enterprises-do-before-scaling-agentic-ai-in-production/
[108] Stellar Cyber. (n.d.). Top agentic AI security threats in late 2026. Available at: https://stellarcyber.ai/learn/agentic-ai-securiry-threats/
[109] Anbiaee, Z., Rabbani, M., Mirani, M., Piya, G., Opushnyev, I., Ghorbani, A., & Dadkhah, S. (2026). Security threat modeling for emerging AI-agent protocols: A comparative analysis of MCP, A2A, Agora, and ANP. *arXiv preprint arXiv:2602.11327*.
[110] Y. Bengio, S. Mindermann, et al. (2025). International AI Safety Report. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
[111] Interview IDs IT_8, IT_13, IT_15, as well as other experts consulted in the study.
[112] Findings of stakeholder consultations carried out within the project..
[113] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[114] Interview ID IT_2. (2026)
[115] ITU. (2025). The Annual AI Governance Report 2025: Steering the Future of AI. Available at: https://www.itu.int/epublications/en/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai/en
[116] Zhuo, T. et al. (2026). To Defend Against Cyber Attacks, We Must Teach AI Agents to Hack. arXiv preprint arXiv:2602.02595.
[117] Interview ID IT_13. Also: SailPoint. (2025). AI agents: The new attack surface. Available at: https://www.sailpoint.com/en-au/identity-library/ai-agents-attack-surface
[118] ITU. (2025). The Annual AI Governance Report 2025: Steering the Future of AI. Available at: https://www.itu.int/epublications/en/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai/en
[119] Liu, Y., Zhang, R., Luo, H., Lin, Y., Sun, G., Niyato, D., Du, H., Xiong, Z., Wen, Y., Jamalipour, A. and Kim, D.I. (2025). Secure multi-LLM agentic AI and agentification for edge general intelligence by zero-trust: A survey. arXiv preprint arXiv:2508.19870.
[120] Defined as "an approach that centres user safety in the design and development of products and services" in Bengio, Y., et al. (2025). International AI safety report 2025. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
[121] Bengio, Y., et al. (2025). International AI safety report 2025. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
[122] Nokia. (2025). Ensure a robust AI-native 6G security posture. Available at: https://www.nokia.com/6g/ensure-a-robust-ai-native-6g-security-posture/
[123] Interview ID IT_2.
[124] Li, X., & Wang, A. (2026). Dynamic multi-agents secured collaboration infrastructure architecture (Internet-Draft draft-li-dmsc-inf-architecture-07). IETF Datatracker. https://datatracker.ietf.org/doc/html/draft-li-dmsc-inf-architecture-07
[125] Interview ID IT_14.
[126] Interview ID IT_6.
[127] See more at: https://hexa-x-ii.eu/objectives/
[128] Interview ID IT_8.
[129] Marcus, G. (2026). AI bot swarms threaten to undermine social media. Marcus on AI (Substack). Available at: https://garymarcus.substack.com/p/ai-bot-swarms-threaten-to-undermine
[130] Bengio, Y., et al. (2025). International AI safety report 2025. Available at: https://internationalaisafetyreport.org/publication/international-ai-safety-report-2025
[131] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[132] Forough, J., Kogias, M., & Haddadi, H. (2026). When Agents Handle Secrets: A Survey of Confidential Computing for Agentic AI. arXiv preprint arXiv:2605.03213.
[133] Barcevičius, E., et al. (2025). Digital identity and identifiers in the era of Web 4.0 and virtual worlds [Thematic paper]. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://op.europa.eu/en/publication-detail/-/publication/4f7b8585-4829-11f1-8095-01aa75ed71a1/language-en
[134] Findings of stakeholder consultations carried out within the project.
[135] Interview ID IT_18
[136] Infocomm Media Development Authority. (2026). Model AI governance framework for agentic AI (Version 1.5). https://www.imda.gov.sg
[137] European Commission. (2026). Action Plan on Cybersecurity and Artificial Intelligence. COM(2026) 577 final. https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence
[138] Interview ID IT_18, 15 and 13
[139] Interview ID IT_15.
[140] Findings of stakeholder consultations carried out within the project..

[141] European Commission. (2026). Action Plan on Cybersecurity and Artificial Intelligence. COM(2026) 577 final. https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence
[142] Liu, Y., Li, P., Wei, Z., Xie, C., Hu, X., Xu, X., Zhang, S., Han, X., Yang, H., & Wu, F. (2026). InfiGUIAgent: A Multimodal Generalist GUI Agent with Native Reasoning and Reflection. Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers). Available at: https://aclanthology.org/2026.eacl-long.45/
[143] Sager, P. J., Meyer, B., Yan, P., von Wartburg-Kottler, R., Etaiwi, L., Enayati, A., Nobel, G., Abdulkadir, A., Grewe, B. F., & Stadelmann, T. (2026). A Comprehensive Survey of Agents for Computer Use. Journal of Artificial Intelligence Research. Available at: https://www.jair.org/index.php/jair/article/view/19490
[144] Janakiram, M. (2026, May 17). The Mac mini just became infrastructure. The New Stack. Available at: https://thenewstack.io/mac-mini-agent-infrastructure/
[145] Interview ID IT_14
[146] Janakiram, M. (2026, May 17). The Mac mini just became infrastructure. The New Stack. Available at: https://thenewstack.io/mac-mini-agent-infrastructure/
[147] Cao, T., Lim, B., Liu, Y., Sui, Y., Li, Y., Deng, S., Lu, L., Oo, N., Yan, S., & Hooi, B. (2025). VPI-Bench: Visual Prompt Injection Attacks for Computer-Use Agents. Available at: https://arxiv.org/abs/2506.02456
[148] Warren, T. (2025, November 18). Microsoft is turning Windows into an 'agentic OS,' starting with AI agents in the taskbar. The Verge. Available at: https://www.theverge.com/news/821948/microsoft-windows-11-ai-agents-taskbar-integration
[149] Bowden, Z. (2025, November 18). Agentic OS will change the way Windows works forever. Yahoo Tech. Available at: https://tech.yahoo.com/ai/copilot/articles/agentic-os-change-way-windows-143015005.html
[150] Hoffman, C. (2026). At Build 2026, Microsoft Sent a Clear Message: Copilot+ PCs No Longer Matter. PC Mag. Available at: https://uk.pcmag.com/ai/165412/at-build-2026-microsoft-sent-a-clear-message-copilot-pcs-no-longer-matter
[151] Mehta, I. (2026, May 12). Google brings agentic AI and vibe-coded widgets to Android. TechCrunch. Available at: https://techcrunch.com/2026/05/12/google-brings-agentic-ai-and-vibe-coded-widgets-to-android/
[152] McCullough, M. (2026, February 25). The Intelligent OS: Making AI agents more helpful for Android apps. Android Developers Blog. Available at: https://android-developers.googleblog.com/2026/02/the-intelligent-os-making-ai-agents.html
[153] Seitz, P. (2026, June 5). Apple set to introduce AI-powered Siri digital assistant at WWDC. Investor's Business Daily. Available at: https://www.investors.com/news/technology/apple-stock-new-siri-assistant-wwdc-2026/
[154] Perez, S. (2026, June 4). Apple approves Poke as the first AI agent on its Messages for Business platform. TechCrunch. Available at: https://techcrunch.com/2026/06/04/apple-approves-poke-as-the-first-ai-agent-on-its-messages-for-business-platform/
[155] Virtual Worlds Partnership. (2025, June). Strategic Research & Innovation Agenda: Virtual Worlds – Solving Real World Problems. Available at: https://www.virtualworldsassociation.eu/actions/strategic-research-innovation-agenda-virtual-worlds-eu
[156] Windows Forum. (2026). Windows Recall Privacy Debate: On-Device AI and the ESU Dilemma. Available at: https://windowsforum.com/threads/windows-recall-privacy-debate-on-device-ai-and-the-esu-dilemma.400682/
[157] Arghire, I. (2025). Microsoft Highlights Security Risks Introduced by New Agentic AI Feature. Security Week. Available at: https://www.securityweek.com/microsoft-highlights-security-risks-introduced-by-new-agentic-ai-feature/
[158] Cao, T., Lim, B., Liu, Y., Sui, Y., Li, Y., Deng, S., Lu, L., Oo, N., Yan, S., & Hooi, B. (2025). VPI-Bench: Visual Prompt Injection Attacks for Computer-Use Agents. Available at: https://arxiv.org/abs/2506.02456
[159] Efstathopoulos, P., Koetzle, L., & Pasquini, D. (2026, April 9). Is That a Bad Apple in Your Pocket? We Used Prompt Injection to Hijack Apple Intelligence. RSAC Conference. Available at: https://www.rsaconference.com/library/blog/is-that-a-bad-apple-in-your-pocket-we-used-prompt-injection-to-hijack-apple-intelligence
[160] OWASP Foundation. (2025). AI Agent Security Cheat Sheet. OWASP Cheat Sheet Series. Available at: https://cheatsheetseries.owasp.org/cheatsheets/AI_Agent_Security_Cheat_Sheet.html
[161] Chaduvula, S. et al. (2026). From Features to Actions: Explainability in Traditional and Agentic AI Systems. arXiv:2602.06841. Available at: https://arxiv.org/abs/2602.06841
[162] Pirch, L. et al. (2026). Toward Securing AI Agents Like Operating Systems. arXiv:2605.14932. Available at: https://arxiv.org/html/2605.14932v1
[163] Chung, J. & Badhe, S. (2026). Local Is Not a Sufficient Privacy Boundary: Governing OS-Integrated On-Device AI. arXiv:2606.10173. Available at: https://arxiv.org/html/2606.10173v1
[164] OWASP. (n.d.). LLM06 – Excessive Agency. Available at: https://genai.owasp.org/llmrisk/llm062025-excessive-agency/
[165] BeyondTrust. (n.d.). Using Endpoint Privilege Management with local AI agents. Available at: https://docs.beyondtrust.com/epm-wm/docs/using-bt-endpoint-privilege-management-with-local-ai-agents
[166] Pirch, L. et al. (2026). Toward Securing AI Agents Like Operating Systems. arXiv:2605.14932. Available at: https://arxiv.org/html/2605.14932v1
[167] Chung, J. & Badhe, S. (2026). Local Is Not a Sufficient Privacy Boundary: Governing OS-Integrated On-Device AI. arXiv:2606.10173. Available at: https://arxiv.org/html/2606.10173v1
[168] Arghire, I. (2025). Microsoft Highlights Security Risks Introduced by New Agentic AI Feature. Security Week. Available at: https://www.securityweek.com/microsoft-highlights-security-risks-introduced-by-new-agentic-ai-feature/
[169] European Commission. (2026). AIXPERT: An agentic, multi-layer, GenAI-powered backbone to make an AI system explainable, accountable, and transparent. CORDIS. Available at: https://cordis.europa.eu/project/id/101214389
[170] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf

[171] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[172] Bostoen, F. & Kramer, J. (2025). Is the DMA Ready for Agentic AI? Centre on Regulation in Europe (CERRE). Available at: https://research.tilburguniversity.edu/en/publications/is-the-dma-ready-for-agentic-ai
[173] Ibid.
[174] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[175] Colangelo, G. (2026). The DMA Meets the New Intermediaries: AI Agents and the Future of Gatekeeper Regulation. ICLE White Paper. Available at: https://laweconcenter.org/wp-content/uploads/2026/05/Colangelo-The-DMA-Meets-the-New-Intermediaries.pdf
[176] Stanley, R., Verma, A., Tsai, L., Kallas, K., & Kumar, S. (2026). An AI Agent Execution Environment to Safeguard User Data. Available at: https://arxiv.org/abs/2604.19657
[177] Reimsbach-Kounatze, C., & Ishikawa, S. (2025). Sharing trustworthy AI models with privacy-enhancing technologies. OECD Artificial Intelligence Papers, No. 38. OECD Publishing. Available at: https://www.oecd.org/en/publications/sharing-trustworthy-ai-models-with-privacy-enhancing-technologies_a266160b-en.html
[178] ITU. (2025). The Annual AI Governance Report 2025: Steering the Future of AI. Available at: https://www.itu.int/epublications/en/publication/the-annual-ai-governance-report-2025-steering-the-future-of-ai/en
[179] Linux Foundation. (2025). Linux Foundation Announces the Formation of the Agentic AI Foundation (AAIF), Anchored by New Project Contributions Including Model Context Protocol (MCP), goose and AGENTS.md. Available at: https://www.linuxfoundation.org/press/linux-foundation-announces-the-formation-of-the-agentic-ai-foundation
[180] Barcevičius, E., et al. (2025). Background document: Input for the Global Multistakeholder High-Level Conference on Governance for Web 4.0 and Virtual Worlds, 31 March – 1 April 2025. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://ec.europa.eu/newsroom/dae/redirection/document/113701
[181] Google LLC. (2025). *Agent2Agent (A2A): A new era of agent interoperability*. Google Developers Blog. https://developers.googleblog.com/en/a2a-a-new-era-of-agent-interoperability/
[182] Proser, Z. (2025). Agent to agent, not tool to tool: an engineer's guide to Google's A2A protocol. WorkOS. Available at: https://workos.com/blog/engineers-guide-to-googles-a2a-protocol
[183] Ecma International. (2026). TC56. Available at: https://ecma-international.org/technical-committees/tc56/
[184] W3C. (n.d.). AI Agent Protocol Community Group. Available at: https://www.w3.org/groups/cg/agentprotocol/
[185] W3C. (n.d.). Agent Trust Protocol (ATP) Community Group. Available at: https://www.w3.org/community/atp/
[186] W3C. (n.d.). Agent Identity Registry Protocol Community Group. Available at: https://www.w3.org/community/agent-identity/
[187] Du, C., Wang, C., Chao, Y., Xie, X., & Cui, Y. (2025). AI agent communication from internet architecture perspective: Challenges and opportunities. arXiv. Available at: https://arxiv.org/pdf/2509.02317
[188] Interview ID IT_13; also Sharma, R., de Vos, M., Chari, P., Raskar, R., & Kermarrec, A. M. (2025). Collaborative agentic AI needs interoperability across ecosystems. *arXiv preprint arXiv:2505.21550*
[189] Chen, W. (2026). "Headless"? What is it and do I need to go there? Infoblox Blog. https://www.infoblox.com/blog/security/headless-what-is-it-and-do-i-need-to-go-there/
[190] Interview ID IT_14.
[191] Zhang, H. (2026). Security Analysis of Multi-agents Secured Communication and Limitations of Existing Protocols. DMSC Working Group. IETF. Available at: https://datatracker.ietf.org/doc/html/draft-zhang-dmsc-mas-communication-00
[192] Dunn, J. (2026). Open source maintainers are being targeted by AI agent as part of 'reputation farming'. InfoWorld. Available at: https://www.infoworld.com/article/4132851/open-source-maintainers-are-being-targeted-by-ai-agent-as-part-of-reputation-farming.html
[193] Interview ID IT_13.
[194] Interview IDs IT_1, IT_13, IT_15, IT_18.
[195] Forthcoming paper "Post-Quantum Cryptography" to be published within the WEB4HUB project.
[196] Allani, S., Ben Yahia, S., & Esposito, C. (2026). A sovereignty-aware and auditable protocol for post-quantum multi-agent communication [Working paper]. SSRN. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6900312
[197] Barcevičius, E., et al. (2025). Background document: Input for the Global Multistakeholder High-Level Conference on Governance for Web 4.0 and Virtual Worlds, 31 March – 1 April 2025. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://ec.europa.eu/newsroom/dae/redirection/document/113701 ; Barcevičius, E., et al. (2025). Digital identity and identifiers in the era of Web 4.0 and virtual worlds [Thematic paper]. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://op.europa.eu/en/publication-detail/-/publication/4f7b8585-4829-11f1-8095-01aa75ed71a1/language-en
[198] Interview ID IT_13; also Sharma, R., de Vos, M., Chari, P., Raskar, R., & Kermarrec, A. M. (2025). Collaborative agentic AI needs interoperability across ecosystems. *arXiv preprint arXiv:2505.21550*.
[199] Arkko, J. and Hardie, T. (2021). RFC 8980 - Report from the IAB Workshop on Design Expectations vs. Deployment Reality in Protocol Development. IETF. Available at: https://datatracker.ietf.org/doc/rfc8980/
[200] Kuerbis, B., & Mueller, M. (2020). The hidden standards war: economic factors affecting IPv6 deployment. *Digital Policy, Regulation and Governance*, *22*(4), 333-361.
[201] See more at: https://nostr.com/
[202] Interview ID IT_12.
[203] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[204] Interview ID IT_15.
[205] Interview ID IT_12.
[206] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.

[207] Nhi Management Group. (2026). Agentic AI Regulation in the EU and UK Exposes Governance Gaps. Available at: https://nhimg.org/articles/agentic-ai-regulation-in-the-eu-and-uk-exposes-governance-gaps/
[208] The Future Society. (2025). Ahead of the Curve: Governing AI Agents under the EU AI Act. Available at: https://thefuturesociety.org/how-ai-agents-are-governed-under-the-eu-ai-act/
[209] Nannini, L., Smith, A. L., Maggini, M. J., Panai, E., Feliciano, S., et al. (2026). Ai agents under EU law. *arXiv preprint arXiv:2604.04604*.
[210] Interview ID IT_15.
[211] China, C. and Goodwin, M. (2025). What is DNS (Domain Name System)? IBM. Available at: https://www.ibm.com/think/topics/dns
[212] Guo, S., et al. (2026). Agent Discovery in Internet of Agents: Challenges and Solutions. *IEEE Network*.
[213] Mozley, J., et al. (2026). Brokered Agent Network for DNS AI Discovery. Available at: https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-bandaid-00.html
[214] Barcevičius, E., et al. (2025). Digital identity and identifiers in the era of Web 4.0 and virtual worlds [Thematic paper]. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://op.europa.eu/en/publication-detail/-/publication/4f7b8585-4829-11f1-8095-01aa75ed71a1/language-en
[215] Interview ID IT_14.
[216] Mozley, J. et al. (2026). Brokered Agent Network for DNS AI Discovery. Available at: https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-bandaid-00.html
[217] Guo, S., et al. . (2026). Agent Discovery in Internet of Agents: Challenges and Solutions. *IEEE Network*.
[218] Infoblox. (2026). Agent Discovery: A Foundational Security Issue for the Agentic Web. Available at: https://www.infoblox.com/blog/company/agent-discovery-a-foundational-security-issue-for-the-agentic-web/
[219] Guo, S., et al. (2026). Agent Discovery in Internet of Agents: Challenges and Solutions. *IEEE Network*.
[220] Guo, S., et al. . (2026). Agent Discovery in Internet of Agents: Challenges and Solutions. *IEEE Network*.
[221] Zhang, H. (2026). Security Analysis of Multi-agents Secured Communication and Limitations of Existing Protocols. DMSC Working Group. IETF. Available at: https://datatracker.ietf.org/doc/html/draft-zhang-dmsc-mas-communication-00
[222] Zylos. (2026). Agent-to-Agent Interoperability Protocols: A2A, ACP, and ANP in Production. Available at: https://zylos.ai/research/2026-04-18-agent-to-agent-interoperability-protocols/
[223] Bu, J. and Krishnan, S. (2026). Announcing the Agentic Resource Discovery specification. Google. Available at: https://developers.googleblog.com/announcing-the-agentic-resource-discovery-specification/
[224] Gabalina, R., Yevdokymova, O. and Zepcan, A. (Forthcoming). Decentralised data and service architectures in the era of Web 4.0 and virtual worlds: Thematic paper. PPMI for the European Commission
[225] Interview ID IT_14.
[226] Interview ID IT_14.
[227] Mozley, J. et al. (2026). Brokered Agent Network for DNS AI Discovery. Available at: https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-bandaid-00.html
[228] DNSimple. (n.d.). The DNSSEC chain of trust. DNSimple Help. https://support.dnsimple.com/articles/dnssec-chain-of-trust/
[229] Hoffman, P., & Schlyter, J. (2012). The DNS-Based Authentication of Named Entities (DANE) Transport Layer Security (TLS) protocol: TLSA (RFC 6698). Internet Engineering Task Force. https://datatracker.ietf.org/doc/rfc6698/
[230] Interview ID IT_14.
[231] The Linux Foundation. (2026). Linux Foundation announces DNS-AID project to advance decentralized AI agent discovery [Press release]. https://www.linuxfoundation.org/press/linux-foundation-announces-dns-aid-project-to-advance-decentralized-ai-agent-discovery
[232] Mozley, J. et al. (2026). Brokered Agent Network for DNS AI Discovery. Available at: https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-bandaid-00.html
[233] Interview ID IT_14.
[234] Infoblox. (2026). Agent Discovery: A Foundational Security Issue for the Agentic Web. Available at: https://www.infoblox.com/blog/company/agent-discovery-a-foundational-security-issue-for-the-agentic-web/
[235] Interview ID IT_15.
[236] Interview ID IT_13.
[237] Interview ID IT_14.
[238] Ibid.
[239] Parham, A. (2026, June 6). AI agent discovery gets open DNS standard: DNS-AID launches under Linux Foundation. Tech Times. https://www.techtimes.com/articles/317904/20260606/ai-agent-discovery-gets-open-dns-standard-dns-aid-launches-under-linux-foundation.htm
[240] Interview ID IT_15.
[241] Interview ID IT_12.
[242] Ibid.
[243] Huang, K., Narajala, V. S., Habler, I., & Sheriff, A. (2026, February). Agent name service (ans): A universal directory for secure ai agent discovery and interoperability. In *2026 IEEE 5th International Conference on AI in Cybersecurity (ICAIC)* (pp. 1-9). IEEE.
[244] Courtney, S. et al. (2026). Agent Name Service v2 (ANS): A Domain-Anchored Trust Layer for Autonomous AI Agent Identity. Available at: https://datatracker.ietf.org/doc/html/draft-narajala-courtney-ansv2-01

[245] Mittal, A., & De La Cruz, E. (2026). Agent Name Service (ANS): A Proof-of-Concept Trust Layer for Secure AI Agent Discovery, Identity, and Governance in Kubernetes. *arXiv preprint arXiv:2604.26997*
[246] Infoblox, & GoDaddy. (2026). Infoblox and GoDaddy support open standards for AI agent discovery, identity and verification. https://aboutus.godaddy.net/newsroom/news-releases/press-release-details/2026/Infoblox-and-GoDaddy-Support-Open-Standards-for-AI-Agent-Discovery-Identity-and-Verification-2026-0acKMntvOC/default.aspx
[247] Interview ID IT_14
[248] Raskar, R., et al. (2025). Beyond DNS: Unlocking the internet of ai agents via the NANDA index and verified agentfacts. *arXiv preprint arXiv:2507.14263*.
[249] Interview ID IT_14
[250] Ibid.
[251] Bu, J. and Krishnan, S. (2026). Announcing the Agentic Resource Discovery specification. Google. Available at: https://developers.googleblog.com/announcing-the-agentic-resource-discovery-specification/
[252] Interview ID IT_14.
[253] Microsoft Build. (2026). What is Microsoft Entra Agent ID? Availble at: https://learn.microsoft.com/en-us/entra/agent-id/what-is-microsoft-entra-agent-id
[254] Mozley, J., et al. (2026). Brokered Agent Network for DNS AI Discovery. Available at: https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-bandaid-00.html
[255] Interview ID IT_14.
[256] See more at: https://registry.modelcontextprotocol.io/
[257] Ehtesham, A., Singh, A., Gupta, G. K., & Kumar, S. (2025). A survey of agent interoperability protocols: Model context protocol (mcp), agent communication protocol (acp), agent-to-agent protocol (a2a), and agent network protocol (anp). arXiv preprint arXiv:2505.02279.
[258] Singh, A., Ehtesham, A., Lambe, M., Grogan, J., Singh, A., Kumar, S., Muscariello, L., Pandey, V., Sauvage De Saint Marc, G., Chari, P., & Raskar, R. (2025). Evolution of AI agent registry solutions: Centralized, enterprise, and distributed approaches. arXiv preprint arXiv:2508.03095v3. https://arxiv.org/abs/2508.03095
[259] Interview ID IT_14.
[260] Interview ID IT_14.
[261] Li, X., & Wang, A. (2026). Dynamic multi-agents secured collaboration infrastructure architecture (Internet-Draft draft-li-dmsc-inf-architecture-07). Internet Engineering Task Force. https://datatracker.ietf.org/doc/html/draft-li-dmsc-inf-architecture-07
[262] Interview ID IT_14.
[263] Interview ID IT_13.
[264] Interview ID IT_15.
[265] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[266] Infoblox. (2026). Agent Discovery: A Foundational Security Issue for the Agentic Web. Available at: https://www.infoblox.com/blog/company/agent-discovery-a-foundational-security-issue-for-the-agentic-web/
[267] EUR-Lex. (2022). Directive (EU) 2022/2555 of the European Parliament and of the Council of 14 December 2022 on measures for a high common level of cybersecurity across the Union, amending Regulation (EU) No 910/2014 and Directive (EU) 2018/1972, and repealing Directive (EU) 2016/1148 (NIS 2 Directive). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2555
[268] EUR-Lex. (2022). Directive (EU) 2022/2557 of the European Parliament and of the Council of 14 December 2022 on the resilience of critical entities and repealing Council Directive 2008/114/EC (CER Directive). Available at: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32022L2557
[269] Interview ID IT_14.
[270] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[271] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[272] Barcevičius, E., et al. (2025). Digital identity and identifiers in the era of Web 4.0 and virtual worlds [Thematic paper]. European Commission, PPMI, TNO, & WEB4HUB. https://op.europa.eu/en/publication-detail/-/publication/4f7b8585-4829-11f1-8095-01aa75ed71a1/language-en
[273] Gabalina, R., Yevdokymova, O. and Zepcan, A. (Forthcoming). Agentic AI: State of Play. PPMI for the European Commission.
[274] Greenblatt, R., et al. (2024). Alignment faking in large language models. arXiv preprint arXiv:2412.14093.
[275] Koorndijk, J. (2025). Empirical Evidence for Alignment Faking in a Small LLM and Prompt-Based Mitigation Techniques. In Proceedings of the AAAI Symposium Series (Vol. 7, No. 1, pp. 198-205).
[276] Findings of stakeholder consultations carried out within the project.
[277] Kokotajlo, D., et al. (2025). AI 2027. Available at: https://ai-2027.com
[278] AlSayyad, A., Huang, K. Y., & Pal, R. (2026). AgentTrace: A Structured Logging Framework for Agent System Observability. In LLM-based Multi-Agent Systems: Towards Responsible, Reliable, and Scalable Agentic Systems. https://arxiv.org/abs/2602.10133
[279] Errico, H. (2026). Autonomous Action Runtime Management (AARM): A System Specification for Securing AI-Driven Actions at Runtime. arXiv preprint arXiv:2602.09433.
[280] Lupinacci, M., et al. (2025). The dark side of llms: Agent-based attacks for complete computer takeover. arXiv preprint arXiv:2507.06850.
[281] Information Commissioner's Office. (2026). ICO tech futures: Agentic AI. Available at: https://ico.org.uk/about-the-ico/research-reports-impact-and-evaluation/research-and-reports/technology-and-innovation/tech-horizons-and-ico-tech-futures/ico-tech-futures-agentic-ai/

[282] European Commission. (2024). Article 14 – Human oversight. Artificial Intelligence Act explainer. https://artificialintelligenceact.eu/article/14/
[283] Interview ID IT_15
[284] National Institute of Standards and Technology. (2026). Announcing the "AI Agent Standards Initiative" for interoperable and secure innovation. https://www.nist.gov/news-events/news/2026/02/announcing-ai-agent-standards-initiative-interoperable-and-secure
[285] Booth, H., Fisher, B., Galluzzo, R., & Roberts, J. (2026). Accelerating the adoption of software and AI agent identity and authorization [Concept paper]. National Institute of Standards and Technology, National Cybersecurity Center of Excellence. https://www.nccoe.nist.gov/projects/software-and-ai-agent-identity-and-authorization
[286] Non-Human Identity Mgmt Group Editorial Team. (2026). Deepfake KYC fraud exposes the limits of selfie-to-ID checks. NHI Management Group. https://nhimg.org/articles/deepfake-kyc-fraud-exposes-the-limits-of-selfie-to-id-checks/
[287] Interview ID ID_10
[288] Interview ID IT_12
[289] Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., Stosz, C., Sastry, G., Cooper, A.F., Balston, G., Adler, S. and Casper, S., (2026). Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies. arXiv preprint arXiv:2601.11699.
[290] ConsentKeys. (n.d.). Privacy-first identity verification. Available at: https://consentkeys.com
[291] Zyphe. (2025). Zero-knowledge proof in KYC: How ZK verification works. Zyphe Resources. https://www.zyphe.com/resources/blog/what-is-zero-knowledge-proof-in-kyc-verification
[292] Spherity GmbH. (2026). Spherity: Decentralized identity for enterprises. https://www.spherity.com/
[293] European Commission. (2026). Blueprint for an age verification solution to help protect minors online. European Commission. https://digital-strategy.ec.europa.eu/en/factpages/blueprint-age-verification-solution-help-protect-minors-online
[294] European Commission. (2026). EU Digital Identity Wallet home. European Commission. https://ec.europa.eu/digital-building-blocks/sites/spaces/EUDIGITALIDENTITYWALLET/pages/694487738/EU+Digital+Identity+Wallet+Home
[295] Chan, A., et al. (2024). Visibility into AI Agents. In The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 3–6, 2024, Rio de Janeiro, Brazil. ACM, 1–16. Page 963.
[296] GlobalData. (n.d.). Trusting AI agents should mean putting them on rails, not letting them run free. Available at: https://www.globaldata.com/newsletter/details/trusting-ai-agents-should-mean-putting-them-on-rails-not-letting-them-run-free_398817/
[297] PYMNTS. (2026). OpenAI says AI agents need to be managed like humans. Available at: https://www.pymnts.com/artificial-intelligence-2/2026/openai-says-ai-agents-need-to-be-managed-like-humans/
[298] Barros, S. (2025). Trusted Identities for AI Agents: Leveraging Telco-Hosted eSIM Infrastructure. arXiv preprint arXiv:2504.16108.
[299] Gallagher, N., & Armstrong, M. M. (2025). What is a digital twin? IBM. Available at: https://www.ibm.com/think/topics/digital-twin
[300] Brundage, M., Dreksler, N., Homewood, A., McGregor, S., Paskov, P., Stosz, C., Sastry, G., Cooper, A.F., Balston, G., Adler, S. and Casper, S., (2026). Frontier AI Auditing: Toward Rigorous Third-Party Assessment of Safety and Security Practices at Leading AI Companies. arXiv preprint arXiv:2601.11699.
[301] Interview ID IT_14
[302] Nordic Institute for Interoperability Solutions. (n.d.). X-Road: Open-source software for unified and secure data exchange. https://x-road.global/
[303] Government of Estonia. (2026, June 17). Prime Minister Michal: Estonia to become first country to create digital identities for AI agents [Press release]. https://www.valitsus.ee/en/news/prime-minister-michal-estonia-become-first-country-create-digital-identities-ai-agents
[304] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[305] Barros, S. (2025). Proof of Humanity: A Multi-Layer Network Framework for Certifying Human-Originated Content in an AI-Dominated Internet. arXiv preprint arXiv:2504.03752.
[306] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[307] Raskar, R., Chari, P., Grogan, J. J., Lambe, M., Lincourt, R., Bala, R., Joshi, A., Singh, A., Chopra, A., Ranjan, R., Gupta, S., Stripelis, D., Gorskikh, M., & Wang, S. (2025). Upgrade or Switch: Do We Need a Next-Gen Trusted Architecture for the Internet of AI Agents? Available at: https://arxiv.org/abs/2506.12003
[308] Bandara, E., Gore, R., Mukkamala, R., Gunaratna, A., Bouk, S.H., Liang, X., Foytik, P., Rahman, A., Rajapakse, S., Kularathna, I. and Karunarathna, P., 2026. Towards an Agent-First Web: Redesigning the Web for AI Agents. arXiv preprint arXiv:2606.19116.
[309] Interview ID IT_14
[310] Ibid.
[311] Ibid.
[312] Findings of stakeholder consultations carried out within the project.
[313] Gabalina, R., Yevdokymova, O. and Zepcan, A. (Forthcoming). Agentic AI: State of Play. PPMI for the European Commission.
[314] Arcep. (2026). Generative AI: Challenges for the Future of the Open Internet. Autorité de régulation des communications électroniques, des postes et de la distribution de la presse. Available at: https://www.arcep.fr/uploads/tx_gspublication/report-generative-ai-challenges-open-internet-january2026.pdf
[315] ibid.
[316] artificialintelligenceact.eu. (2026.). Article 50: Transparency obligations for providers and deployers of AI systems. Retrieved June 28, 2026, from https://artificialintelligenceact.eu/article/50/

[317] European Commission: Directorate-General for Communications Networks, Content and Technology & Puccetti, G. (2026). Technical solutions for marking and detecting AI generated text content in the context of Article 50(2) AI Act, Publications Office of the European Union. https://data.europa.eu/doi/10.2759/7579127
[318] Coalition for Content Provenance and Authenticity. (n.d.). C2PA – Content Credentials and provenance standards. https://c2pa.org/
[319] Association for Information Science and Technology. (2025). ISO/DIS 22144, Authenticity of information – Content credentials. https://www.asist.org/2025/03/19/iso-22144-authenticity-information-standards/
[320] Joint Photographic Experts Group. (n.d.). JPEG Trust (ISO/IEC 21617) – Framework for establishing trust in media. https://jpeg.org/jpegtrust/
[321] European Commission. (2026). Code of Practice on marking and labelling of AI-generated content. https://digital-strategy.ec.europa.eu/en/news/commission-publishes-code-practice-marking-and-labelling-ai-generated-content
[322] Interview ID IT_15.
[323] Ibid.
[324] Nemecek, A., He, H., Cheng, G., & Ayday, E. (2026). Authenticated contradictions from desynchronized provenance and watermarking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 10738-10748).
[325] Cao, J., Li, Q., Zhang, Z., Ni, J., & Lu, R. (2025). Secure and robust watermarking for ai-generated images: A comprehensive survey. arXiv preprint arXiv:2510.02384.
[326] Google DeepMind. (2024). SynthID: Identifying AI-generated content. Google DeepMind. https://deepmind.google/models/synthid/
[327] Bulychev, M., Marchant, N. G., & Rubinstein, B. I. (2026). Watermarks Attack Watermarks: Re-Watermarking as a Generic Removal Strategy. arXiv preprint arXiv:2605.16796.
[328] An, L., Liu, Y., Liu, Y., Zhang, Y., Bu, Y., & Chang, S. (2025). Defending llm watermarking against spoofing attacks with contrastive representation learning. arXiv preprint arXiv:2504.06575.
[329] Alzoubi, Y. I., Al-Ahmad, A., & Kahtan, H. (2022). Blockchain technology as a Fog computing security and privacy solution: An overview. Computer Communications, 182, 129-152.
[330] Open Source Security Foundation. (2025). Model Signing Specification [Software specification]. GitHub. https://github.com/ossf/model-signing-spec
[331] Interview ID_IT_2.
[332] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[333] Mozley, B. & Williams, M. (2026). DNS for AI Agent Discovery (DNS-AID) (draft-mozleywilliams-dnsop-dnsaid-02). Internet Engineering Task Force. https://www.ietf.org/archive/id/draft-mozleywilliams-dnsop-dnsaid-02.html
[334] European Commission: Directorate-General for Communications Networks, Content and Technology & Puccetti, G. (2026). Technical solutions for marking and detecting AI generated text content in the context of Article 50(2) AI Act, Publications Office of the European Union. https://data.europa.eu/doi/10.2759/7579127
[335] Ibid.
[336] Shumailov, I., Shumaylov, Z., Zhao, Y., Papernot, N., Anderson, R., & Gal, Y. (2024). AI models collapse when trained on recursively generated data. Nature, 631(8022), 755-759.
[337] Interview ID IT_4.
[338] Interview ID IT_7.
[339] Interview ID IT_13 & IT_15.
[340] Yu, H., Kim, D., & Kim, Y. B. (2026, April). Retrieval collapses when AI pollutes the web. In Proceedings of the ACM Web Conference 2026 (pp. 8745-8748). https://dl.acm.org/doi/abs/10.1145/3774904.3792955
[341] Noroozian, A., Aldana, L., Arisi, M., Asghari, H., Avila, R., Bizzaro, P. G., ... & Wasielewski, A. (2025). Generative AI and the Future of the Digital Commons: Five Open Questions and Knowledge Gaps. arXiv preprint arXiv:2508.06470.
[342] European Commission. (2025). Communication on European tech sovereignty, accompanied by an EU Open Source Strategy. https://digital-strategy.ec.europa.eu/en/library/communication-european-tech-sovereignty-accompanied-eu-open-source-strategy
[343] Interview ID IT_15.
[344] Interview ID IT_14.
[345] PYMNTS. (2025). Shopify president: Agentic AI drives 11x spike in orders. PYMNTS.com. https://www.pymnts.com/news/ecommerce/2025/shopify-president-agentic-ai-drives-11x-spike-in-orders
[346] Ibid.
[347] Kaushik, N. (2026). Building the trust layer for agentic payments with AP2 and Verifiable Intent. FIDO Alliance. https://fidoalliance.org/building-the-trust-layer-for-agentic-payments-with-ap2-and-verifiable-intent/
[348] Vaziry, A., Garzon, S. R., & Küpper, A. (2025). Towards multi-agent economies: Enhancing the a2a protocol with ledger-anchored identities and x402 micropayments for ai agents. arXiv preprint arXiv:2507.19550.
[349] Vaziry, A., Garzon, S. R., & Küpper, A. (2025). Towards multi-agent economies: Enhancing the a2a protocol with ledger-anchored identities and x402 micropayments for ai agents. arXiv preprint arXiv:2507.19550.
[350] Ibid.
[351] Accenture. (2026). Agentic commerce and the future of payments. Accenture Banking Blog. https://www.accenture.com/us-en/blogs/banking/agentic-commerce-payments

[352] Boston Consulting Group. (2025). Global Payments Report 2025: The Future Is (Anything but) Stable. https://www.bcg.com/publications/2025/global-payments-transformation-amid-instability
[353] Davidovic, S., & Tourpe, H. (2026, April). How agentic AI will reshape payments (IMF Note 2026/004). International Monetary Fund. https://www.elibrary.imf.org/view/journals/068/2026/004/article-A001-en.xml
[354] Vaziry, A., Garzon, S. R., & Küpper, A. (2025). Towards multi-agent economies: Enhancing the A2A protocol with ledger-anchored identities and x402 micropayments for AI agents. arXiv preprint arXiv:2507.19550.
[355] Accenture. (2026, May 26). Agentic commerce and the future of payments. Accenture Banking Blog.
[356] Boston Consulting Group. (2025). Global Payments Report 2025: The Future Is (Anything but) Stable. https://www.bcg.com/publications/2025/global-payments-transformation-amid-instability
[357] Ibid.
[358] Interview ID IT_12
[359] Ibid.
[360] Davidovic, S., & Tourpe, H. (2026, April). How agentic AI will reshape payments (IMF Note 2026/004). International Monetary Fund. https://www.elibrary.imf.org/view/journals/068/2026/004/article-A001-en.xml
[361] Ibid.
[362] Interview ID IT_13
[363] Ibid.
[364] Interview ID IT_14
[365] Chen, W. (2026). "Headless"? What is it and do I need to go there? Infoblox Blog. https://www.infoblox.com/blog/security/headless-what-is-it-and-do-i-need-to-go-there/
[366] Barcevičius, E., et al. (2025). Digital identity and identifiers in the era of Web 4.0 and virtual worlds [Thematic paper]. European Commission, PPMI, TNO, & WEB4HUB. Available at: https://op.europa.eu/en/publication-detail/-/publication/4f7b8585-4829-11f1-8095-01aa75ed71a1/language-en
[367] Sitouah, N., Esposito, M., & Bruschi, F. (2026). Self-Sovereign Identity and eIDAS 2.0: An Analysis of Control, Privacy, and Legal Implications. arXiv preprint arXiv:2601.19837. https://arxiv.org/pdf/2601.19837
[368] South, T., et al. (2025). Authenticated delegation and authorized AI agents. arXiv:2501.09674. https://arxiv.org/abs/2501.09674
[369] Government of Estonia. (2026, June 17). Prime Minister Michal: Estonia to become first country to create digital identities for AI agents [Press release]. https://www.valitsus.ee/en/news/prime-minister-michal-estonia-become-first-country-create-digital-identities-ai-agents
[370] x402 Foundation. (n.d.). x402 — Payment Required: Internet-native payments standard. https://www.x402.org/
[371] Fewsats, Inc. (n.d.). L402: An HTTP-based payment flow for MCP, AI agents, and more. https://www.l402.org/
[372] Business Wire. (2026). Alipay AI Pay launches new service enabling OpenClaw-type AI agents to make payments [Press release]. https://www.businesswire.com/news/home/20260421171651/en/Alipay-AI-Pay-Launches-New-Service-Enabling-OpenClaw-type-AI-Agents-to-Make-Payments.
[373] Klarna. (2025). Klarna launches Agentic Product Protocol: The open standard that makes 100M+ products instantly discoverable by AI agents [Press release]. https://www.klarna.com/international/press/klarna-launches-agentic-product-protocol-the-open-standard-that-makes-100m/
[374] VerityScore. (2026). Schema.org Product for Shopify: Required fields & common mistakes. VerityScore Knowledge Base. https://verityscore.io/en/kb/schema-org/
[375] GS1. (n.d.). GS1 style guide (Current standard). https://www.gs1.org/standards/gs1-style-guide/current-standard
[376] Chaffer, T. J., von Goins II, C., Cotlage, D., Okusanya, B., & Goldston, J. (2024). Decentralized governance of AI agents. arXiv preprint arXiv:2412.17114. https://arxiv.org/abs/2412.17114
[377] Nuvei. (2026, June 18). Why KYA is the new control point for agentic commerce. Available at: https://www.nuvei.com/posts/know-your-agent-kya-agentic-commerce
[378] Chaffer, T. J., von Goins II, C., Cotlage, D., Okusanya, B., & Goldston, J. (2024). Decentralized governance of AI agents. arXiv:2412.17114
[379] Ibid.
[380] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[381] European Commission. (2026). AI Office publishes frontier AI expert findings on EU competitiveness, sovereignty and security. Available at: https://digital-strategy.ec.europa.eu/en/library/ai-office-publishes-frontier-ai-expert-findings-eu-competitiveness-sovereignty-and-security
[382] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[383] WEB4HUB expert roundtable. (2026, 2 July). Opportunities for Europe in the Web 4.0 and Agentic AI Era.
[384] European Commission. (2026). AI Office publishes frontier AI expert findings on EU competitiveness, sovereignty and security. Available at: https://digital-strategy.ec.europa.eu/en/library/ai-office-publishes-frontier-ai-expert-findings-eu-competitiveness-sovereignty-and-security
[385] European Commission. (2026). Action Plan on Cybersecurity and Artificial Intelligence. COM(2026) 577 final. https://digital-strategy.ec.europa.eu/en/library/eu-action-plan-cybersecurity-and-artificial-intelligence