The Accountable Agents Framework (AAF): A Blueprint for AI Agent IDs

Adam Khoja[1,*,†], **Leo Wu**[1,*,†], **Wayne Chang**[2,†], **Dazza Greenwood**[3,4,†], **Sam Boger**[5,†], **Jakub Kraus**[1,†], **Anita Srinivasan**[6,†], **Ian W Eisenberg**[7,†], **Christian Turner**[8,†], **Ankit Agarwal**[9], **Yonathan A. Arbel**[10,11], **Ben Bariach**[12], **Daniel Carpenter**[13], **Stephen Casper**[13], **Inyoung Cheong**[14], **James Fox**[15], **Rokas Gipiškis**[16,17], **Simon Goldstein**[18], **Thomas Hardjono**[19], **Zoë Hitzig**[20], **Gail Hodges**[21], **Edward Hughes**[22], **Jieun Hwang**[23], **Brian Jabarian**[24], **William Marks**[14,25], **Soham Mehta**[26], **Niloofar Mireshghallah**[24], **Matthew Mittelsteadt**[27], **Evan Miyazono**[28], **Tim O'Reilly**[29,30], **Simon Osindero**[31], **Alex Pentland**[4], **Richard Ren**[1], **Anka Reuel**[4], **Noah Ringler**[32], **Peter Salib**[33,34], **Ido Salomon**[35], **Bruce Schneier**[36], **Christian Schroeder de Witt**[12], **Leon Staufer**[5], **Audrey Tang**[12], **Nenad Tomašev**[31], **Tricia Wang**[37], **Kevin Wei**[38], **Gabriel Weil**[33,34], **E. Glen Weyl**[39], **Jesse Wright**[12], **Roman Yampolskiy**[40], **Jonathan Zittrain**[14], **Dan Hendrycks**[1,‡], **Gillian K. Hadfield**[20,36,41,‡]

[1]Center for AI Safety [2]Spruce Systems [3]Civics.com [4]Stanford University [5]SASH [6]MATS [7]Credo AI [8]University of Georgia [9]Skyfire [10]University of Alabama [11]Center for Law & AI Risk [12]University of Oxford [13]Harvard University [14]Berkman Klein Center, Harvard University [15]Schmidt Sciences [16]AI Standards Lab [17]Vilnius University [18]University of Hong Kong [19]MIT [20]Johns Hopkins University [21]OpenID Foundation [22]Inherent Laboratories [23]Personal Information Protection Commission, Republic of Korea [24]Carnegie Mellon University [25]AfterQuery [26]Foundation for American Innovation [27]Institute for AI Policy and Strategy [28]Atlas Ignota [29]O'Reilly Media [30]AI Disclosures Project [31]Google DeepMind [32]Institute for Security and Technology [33]University of Houston [34]Institute for Law & AI [35]Independent [36]University of Toronto [37]Advanced AI Society [38]Oxford Martin AI Governance Initiative [39]Microsoft Research [40]University of Louisville [41]Vector Institute

[*]Equal contribution [†]Lead authors [‡]Senior authors

Corresponding authors: adamk@safe.ai, leo@safe.ai

Abstract

AI agents are becoming ubiquitous online, but there is no reliable way to identify who is responsible for their actions. This creates an accountability gap and increases hesitation around agent adoption in the economy and society. To help address this challenge, we introduce the Accountable Agents Framework (AAF), a protocol-agnostic framework that uses privacy-preserving IDs to link agents to their principals, such as a legal person. Directories verify principals' identities and issue IDs, allowing agents to make verifiable claims about themselves as they transact. IDs are pseudonymous, revealing nothing about agents on their own, but if an agent causes harm, directories can deanonymize its principal according to an established process. Services can then more safely engage in transactions with agents that present a verified ID, since harm caused by such agents can be more reliably traced back to their principal. This paper specifies requirements an effective agent ID system should meet, and discusses how accountability infrastructure can help integrate agents into society.

1 Introduction

The internet runs on trust [1]. Countless online interactions and services occur between strangers [2]; for most transactions today, trust is nonetheless possible thanks to infrastructure such as encryption,

Note: Listed authors contributed substantive ideas and/or work to the paper. Inclusion as an author does not entail endorsement of all claims in the paper, nor does it imply endorsement by any author's organization.

certificate authorities, and payment protocols. Internet infrastructure enables tens of trillions of dollars of commerce and tremendous cultural value [3–5].

Today, however, a growing share of online activity comes from AI agents: AI systems that act autonomously in the world, usually on behalf of a person or organization [6–8]. Consumers and organizations already delegate consequential tasks to them, from writing code and booking appointments to conducting research and running businesses [9–13]. Cloudflare reported in 2026 that their daily web requests from AI agents had grown by over 1,700% from the prior year [14].

Yet this surge strains an internet designed for humans: high-volume agent traffic can amplify the scale and sophistication of malicious activity, create new opportunities for harm as more affordances are delegated to agents, and complicate assessments of liability if harm does occur [15–17]. These problems are intensified by an accountability gap: there is no standardized way to identify agents or their principals online [8, 18–20].

This accountability gap also impedes agentic commerce [6, 7, 21, 22]. When users grant agents full access to online services, services may struggle to distinguish between intentionally authorized and malicious activity [23]. Some services already block broad categories of automated traffic, precluding beneficial automation and agent access [24–27]. Should we want agents to assist in our economy at machine speed and scale, platforms will need more nuanced options than allowing agents unrestricted access or banning them completely.

To address this accountability gap, we introduce the Accountable Agents Framework (AAF). At the core of the framework is an identifier (*ID*), a *claim* that can link an agent to a principal, such as a legal person, without revealing any information by default. In an AAF-compatible system, principals register themselves and their agents with a *directory*: a trusted entity that verifies principals' identities, retains registration records, and issues IDs and other attestations to agents (Figure 1). As agents transact with counterparties, such as online services, individuals, or other agents, they can present their IDs alongside other verifiable claims about themselves or their principals. Much as license plates allow the accountable use of cars, identifiers for AI agents are a foundational piece of infrastructure that builds the trust needed for agents to benefit society.

Our framework follows a long line of work on agent IDs, agent infrastructure, and digital identity [28–30]. Indeed, systems that fulfill many of the functions we outline here have already been developed or proposed, encompassing decentralized, centralized, and market-driven approaches [31–33]. For more on related work, see Appendix B. This paper specifies the properties an AAF-compatible system should satisfy, establishes a shared language for the components of accountability infrastructure, and examines implementation and governance considerations to guide further development and standardization.

AAF focuses on a necessary first step for accountability infrastructure: technical mechanisms for tracing agents to verified principals. The framework does not prescribe how to structure directories, how to weigh accountability against privacy, how liability is assigned, or when IDs should be required. By leaving these questions open, AAF lets principals, online services, and directories experiment with different arrangements and settle on those with the best trade-offs. As agents proliferate, the implementation of accountability infrastructure will shape how the internet operates.

Section 2 describes the framework, Section 3 discusses implementation, governance, and limitations, and Section 4 concludes. Appendix A provides a detailed specification of the framework, Appendix B discusses related work, Appendix C provides an overview of potential technical implementation options, and Appendix D provides a security analysis.

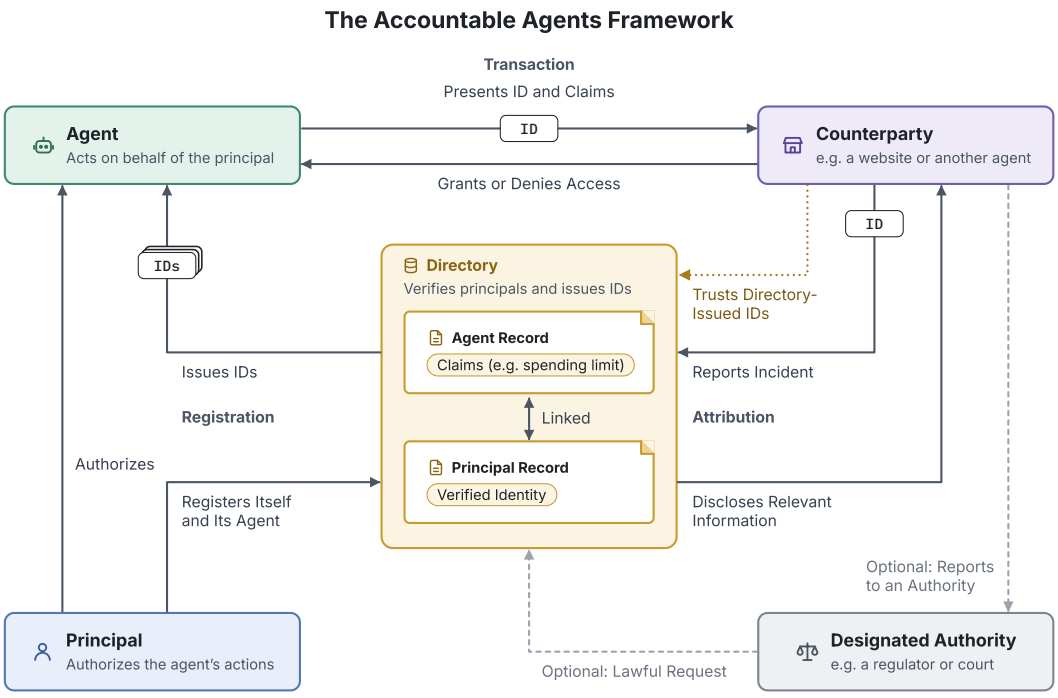


Figure 1: An overview of the Accountable Agents Framework. The framework supports three main categories of interactions. Registration: a principal may register itself and its agent with a directory, which issues IDs to the agent. Transaction: agents present claims (including IDs) to counterparties that can be verified, enabling better decision-making during transactions. Attribution: in case of harm, IDs can also be used to trace agents to their principals and hold them accountable. See Section 2.2 for more detailed analysis of important interactions.

2 The Accountable Agents Framework

This section defines the actors and components of an AAF-compatible system (2.1), how the system's components interact (2.2), and discusses the framework's core advantages (2.3).

2.1 Actors and Components

Principals. In the framework, we define a *principal* as the entity that authorizes an AI agent's activity. The principal may register the agent with a directory, grant it authority, and revoke its registration at any point. Principals are usually legal persons—including specific humans and organizations—but identity can technically also be tied to other entities, such as controllers of web domains, email addresses, or phone numbers.

AI agents. We use the term *agent* to mean AI agents—AI systems that take autonomous actions in the world [34, 35]. Note that this definition differs from the legal concept of agency, which requires a fiduciary relationship between two legal persons [8, 36]. While other work has proposed legal personhood and fiduciary duties for AI agents [37–39], here we do not take a position on these questions.

Since an agent's underlying model and context can change over time, the framework does not distinguish an agent using these characteristics. Agents are identified only by how they reliably authenticate themselves. An agent may authenticate by using methods such as proving possession of cryptographic secrets or network origin-based mechanisms. For the technical purposes of the framework, the agent is defined by its authentication. Agents should therefore support robust authentication, such as securely managing secrets when relying on cryptographic mechanisms [40–42].

Counterparties. *Counterparties* are entities that interact with an agent. Examples include an e-commerce vendor an agent is purchasing from, a person talking to a sales-focused agent, a bank or government portal an agent wants to use, a website an agent is reading or scraping, or another agent. Counterparties decide whether and how to transact with agents, and thus can choose whether and how to use AAF-compatible presentations (defined below) they receive from agents; they can verify agents' claims, request additional information, and/or log presentations. Counterparties can also simply ignore presentations and serve agents indiscriminately, making AAF backward-compatible.

Directories. Directories are entities that enable trust in agent transactions. They hold and verify principals' identities, issue IDs to agents, and attest to agents' claims. A counterparty's trust in directories can thus be extended to the claims an agent presents. Since AAF defines directories by their functional role rather than a specific institutional form, they can be operated by existing identity providers (such as domain registrars or payment networks), dedicated private companies, industry consortia, or public agencies; in Section 3.1, we discuss these design choices and how they balance accountability and privacy.

Beyond requiring directories to retain enough information to link IDs, agents, and principals, AAF is agnostic to the specific data fields that directories collect and to their retention policies. For example, directories might store: agent authorizations; identity proofing records; details about a human principal; the scope of authority and/or intent for the agent; the agent's model, harness, tools, and versions; and financial information. This extensibility allows directories to adapt to new contexts by adopting standards and protocols that are additive to the registration records of principals or agents. Directories and principals can weigh what information should persist based on competing considerations, such as privacy, accountability, cost, and convenience. Counterparties can similarly weigh a directory's characteristics when deciding whether to trust that directory.

Presentations and claims. When an agent contacts a counterparty, it sends an AAF-compatible *presentation*: a set of verifiable claims, including an ID. A claim is a statement about an agent or its principal, such as the AI model behind an agent, the principal's birth year, or the authorizations an agent has. Claims can be attested by the directory or the principal, or self-asserted by the agent; the attesting party is the *issuer* of the claim. For example, an authorization can take the form of a principal-attested claim about what the agent may do. Agents can selectively choose which claims to disclose in presentations, with more sensitive interactions potentially warranting more disclosure.

Counterparties do not need to contact directories to validate individual presentations, just as web browsers can check website certificates without contacting certificate authorities. Using only a presentation and cryptographic materials published by its associated directory, such as the directory's public key, counterparties can verify the validity of claims, including that all claims were indeed signed by their issuers. Agents can authenticate themselves alongside their presentations, so that counterparties can verify that a presentation actually refers to the agent they are interacting with.

Identifiers (IDs). AAF-compatible IDs are special claims, required in each presentation, that allow directories to recover the agent and principal behind a presentation. Like license plates on cars, IDs do not reveal any information about agents or principals to counterparties by default [43]. Unlike standard license plates, IDs in different presentations can also be *unlinkable*, so counterparties cannot associate different presentations with the same agent using the IDs alone. However, these protections do not prevent linkage through disclosed claims or other identifying information.

Alternatively, agents can choose to reuse IDs throughout their interactions with a distinct counterparty or set of counterparties. The framework thus supports persistent identifiers unique to a relationship, also known as pairwise pseudonymous identifiers (*PPIDs*)—with directories attesting they have issued at most one PPID for a given agent and counterparty—to support account continuity, rate-limiting, and reputation (Figure 2). For more discussion on the properties of PPIDs, see Appendix A, PR-8; for commentary on implementation, see Appendix C, Section 7.

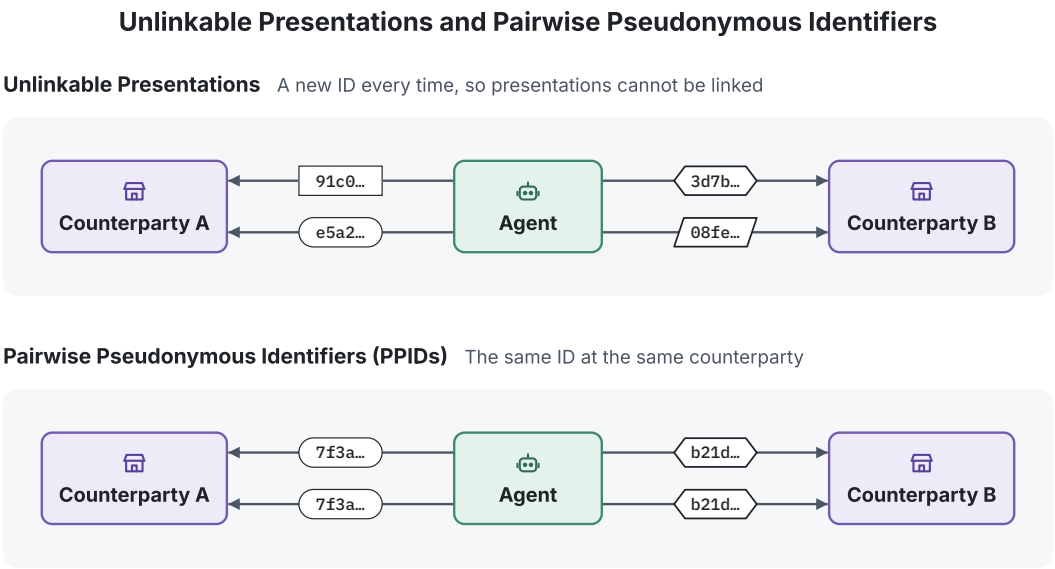


Figure 2: Unlinkable vs. reusable IDs for agents. With reusable IDs (PPIDs), an agent shows each counterparty the same ID every time. IDs can still be unique to counterparties, so counterparties cannot collude to link the agent across services using only their IDs. With unlinkable IDs, each ID is used only once, so even a single counterparty cannot link separate interactions to one agent by looking at the IDs presented.

2.2 Interactions

In this section, we outline specific interactions under AAF. Principals can *register* themselves and their agents in directories; agents can potentially *delegate* by registering sub-agents; principals and directories can *revoke* agents' registrations; principals can *authorize* agent actions; agents can *transact* with counterparties; and directories can assist with *attribution*: recovering the agents and principals behind an ID according to an established process, such as in response to a lawful request.

Principal registration. For directories to confidently attest to a principal's identity and other information, that information must be verified. Thus, directories must engage in *identity proofing* during principal registration up to a certain assurance level, such as NIST's IAL2 standard [44]. To interact with directories for sensitive purposes, such as registering or revoking agents, principals establish their own means of authentication with the directories up to a certain assurance level, such as NIST's AAL2 standard [45]. During registration, all directory information about the principal is included in a *principal record*; the principal's agents can then receive directory-attested claims for any information the directory has verified. Directories perform *Sybil checks* to ensure that a single principal cannot register multiple times [46].

Agent registration. The principal can register an agent with a directory. Like principals, agents must establish their authentication methods with directories during registration and authenticate with the directory to receive IDs. Information the directory receives about an agent, such as its provenance or authorizations, is included in its *agent record*, which is linked to its principal's record. Directories can then attest to the subset of the agent record they can verify. For example, agents can include directory-attested claims about their means of authentication in presentations, such as the agent's public key, so that counterparties can verify that the agent sending a presentation is who it claims to be.

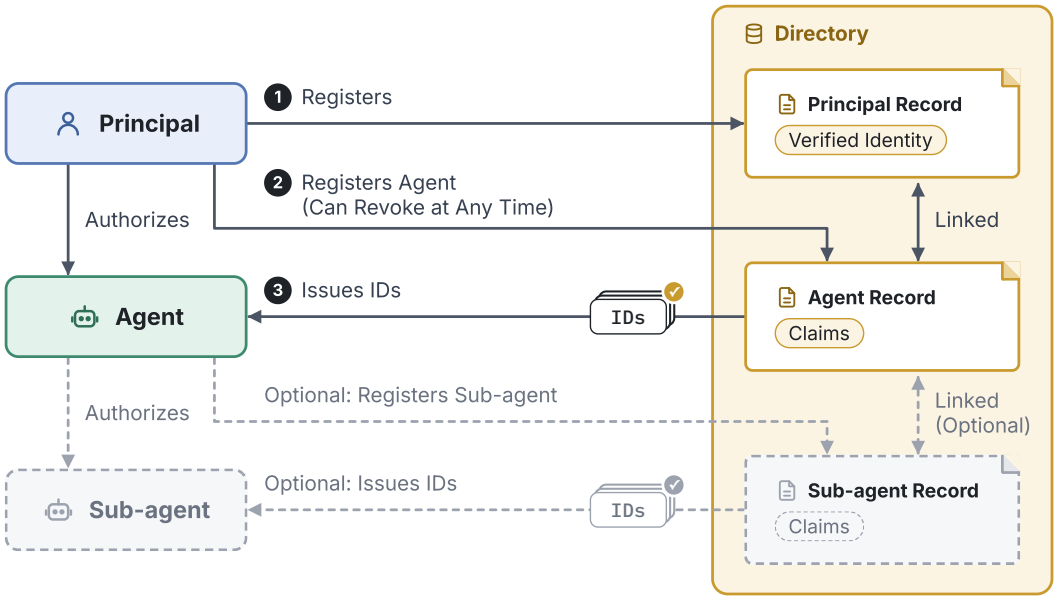


Figure 3: Registration and delegation. A principal registers itself with the directory, which verifies the identity of the principal (1). The principal may then register an agent with the directory (2), which issues IDs for the agent (3). The agent may further delegate to sub-agent(s), acting as a principal and registering sub-agent(s), which receive their own IDs, and so on. The directory maintains a record for each agent and sub-agent and links each agent record to the one above it, such that every chain can be traced back to a verified principal.

Revocation. Principals, as well as directories in some cases, can revoke their attestation of individual claims, or revoke an agent's registration altogether. When counterparties try to validate presentations, the attempt should eventually fail for revoked claims or IDs, similar to how certificate authorities can revoke certificates. Revocation can be warranted when an agent is no longer acting as desired, or in more benign cases when the agent has completed its tasks. See Section 3.5 for additional discussion about revocation.

Agent delegation and multi-agent systems. Directories can potentially record more complex networks of relationships between sets of principals and agents. The implementation complexities and accountability implications of systems with many agents and principals are mostly left out of scope for this paper, though we will briefly discuss them here.

AAF can support *delegating agents* that can register, authorize, and revoke *sub-agents* of their own (see Figure 3). By default, delegation can only narrow an agent's scope. Additionally, AAF is compatible with agents that have multiple principals, such as when different principals authorize different actions, or when agents must act on behalf of many principals simultaneously [47]. Finally, since an agent under AAF is only tied to its means of authentication, it is possible that a team of AI agents may share a single agent record and set of IDs, without individually registering with a directory. In general, upholding accountability becomes more complicated when an outcome involves multiple actors [48–50], but infrastructure that can accurately attribute actions to networks of agents and principals is nonetheless beneficial.

Transactions and authorization. At transaction time, counterparties can request (or agents can proactively send) AAF-compatible presentations. Counterparties can then request additional information, such as authorizations from the agent's principal, and the agent can return with a new presentation with the requested claims. Figure 4 illustrates the steps of a transaction between an agent, counterparty, and authorizing principal. Importantly, principal authorization under AAF is possible without the counterparty learning the principal's identity, as we discuss in Appendix C, Section 4.

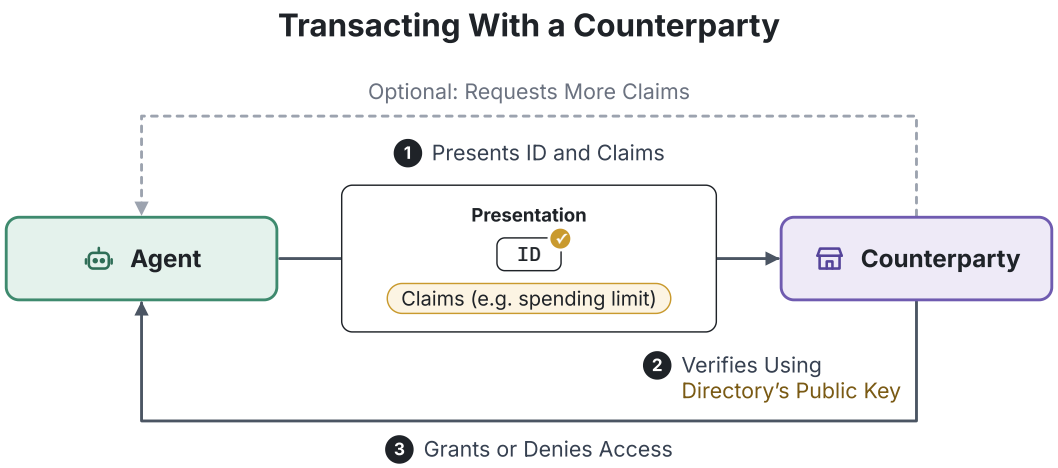


Figure 4: A transaction between an agent and a counterparty. The agent presents certain claims (including the ID) to a counterparty (1). The counterparty verifies the ID and claims, such as through matching against the directory's public key (2). The counterparty may ask the agent for additional claims if required, which the agent can pass on to its principal. Based on information received, the counterparty then determines whether to grant or deny access to the agent (3).

Attribution. AAF-compatible directories support *attribution*: the ability to trace an ID back to the agent and principal records behind it (Figure 5). They do so through *identity escrow* [51], an arrangement that lets the directory recover those records according to an established process, such as with the principal's consent (e.g., for dispute resolution or audits) or in response to a lawful request. Technical methods, such as multi-party cryptography [52], can keep the link between agents and principals private by default and reduce the risk of abuse by directories or authorities; we discuss possible implementations in Appendix C, Section 6, and other trade-offs of our proposal in Section 3.

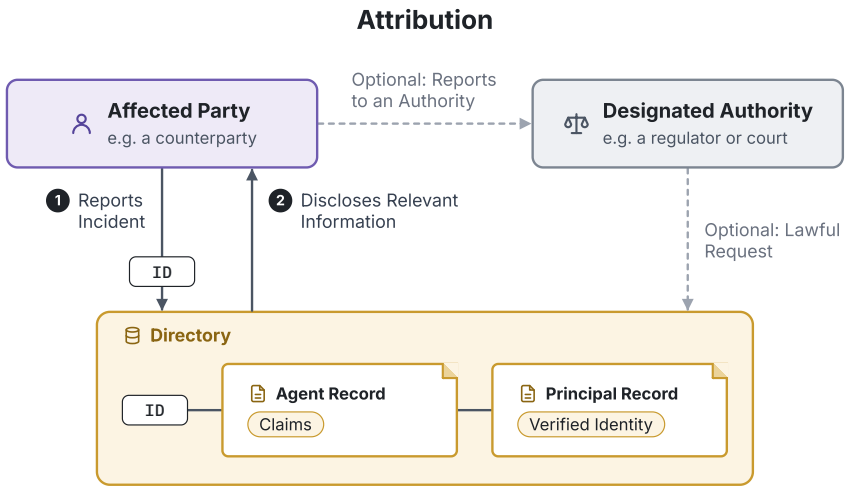


Figure 5: Attribution in cases of harm. The affected party reports an incident and involved IDs to the directory (1). The directory recovers records associated with the ID and discloses relevant information according to an established process (2), enabling attribution and accountability. In some cases, affected parties may ask a designated authority (such as a law enforcement agency, regulator, or court) to submit a lawful request to the directory, for instance to receive a more detailed disclosure.

2.3 Advantages of AAF

This framework has five core advantages over the way agents are identified and trusted today.

Accountable. When an agent causes harm, the agents and principals involved can be identified. Directories can resolve IDs to the agent and principal records behind them through an established process.

Verifiable. Agents can present claims that are attested by directories or principals. A counterparty can verify that each claim was signed by its issuer and not tampered with. Counterparties only need to trust directories, not agents, to trust claims.

Minimally disclosing. Information is shared on a need-to-know basis, which limits counterparties' access to data about an agent and their ability to track its activity across services [53]. The principal's identity does not need to be revealed to counterparties in most cases [54].

Proportionate. Different counterparties can work with different directories based on the degree of assurance they require. A bank or other regulated business may only trust directories with strict security and data compliance practices, while other services may accept weaker forms of assurance.

Interoperable. Directories with compatible practices can be federated. AAF can be minimally implemented using existing digital identity infrastructure, and can extend to new protocols and contexts.

3 Discussion

By establishing trust in agentic transactions, AAF-compatible systems can support the growth of the agentic internet. However, implementing such a framework raises practical governance questions about directory design choices (3.1), adoption dynamics (3.2), how agent IDs help address rogue agents (3.3), security threats to ID systems (3.4), and the framework's limitations (3.5).

3.1 Directory Design Choices

Tracing an agent's actions to a principal requires a reliable link between the two. In AAF, directories establish that link at registration and retain the records needed to recover it. This section explores how directories can fulfill that role, what trade-offs their designs involve, and what precedents they can draw on.

Directories can vary in their centralization and data practices. Directories vary along two main axes: their degree of centralization (from state-operated systems, like China's proposed national agent ID infrastructure, to private companies and industry consortia in technology, finance, and healthcare, to smart contracts on permissionless ledgers such as Ethereum's ERC-8004) and their degree of assurance (from KYC, Legal Entity Identifiers, or NIST's IAL2 standard down to domain control, email, phone, or proof of unique personhood). More centralized, higher-assurance designs offer stronger Sybil resistance and clearer accountability, whereas lightweight or decentralized designs reduce data concentration and onboarding friction. AAF-compatible directories publish their assurance and disclosure practices so principals and counterparties can select designs suited to their sector.

The accountability-privacy trade-off. Providing data to directories entails privacy risks (breaches, insider misuse, commercial tracking, or government overreach), yet accountability infrastructure can enhance privacy on balance by letting counterparties accept minimally disclosing presentations instead of demanding upfront identification or blocking agents outright.

Escrow design and minimal visibility. Even where directories verify principals, they can be designed to be privacy-first: cryptographic ID issuance and revocation methods can keep directories blind to presentation timing and counterparties (Appendix C, Sections 3, 5, and 7). To avoid concentrating trust in a single operator, directories can divide verification, issuance, and record-keeping across separate organizations, or use cryptographic multi-party control requiring independent trustees to cooperate before revealing a principal's identity (Appendix C, Section 6).

3.2 Adoption Dynamics

Market incentives exist to build and operate reliable identification infrastructure. While governments could operate official directories, private directories provide value simply by being a trusted source of information about agents; agent IDs do not need to be widely adopted to provide value. Indeed, most existing agent authentication and identification protocols have been introduced by private companies. Under AAF, directories may be operated by agent providers, counterparties, or principals themselves, while dedicated directory businesses can compete on convenience, latency, availability, and security, creating a form of market-driven governance.

Adoption does not require building directory infrastructure from scratch. The internet already has infrastructure that performs directory-like functions and can directly model or implement AAF-compatible directories. For example, domain registrars require a registrant of record, verify contact channels, and handle abuse or lawful-disclosure requests for redacted records; certificate authorities issue offline-verifiable credentials upon proof of domain control; and payment networks link tokenized transactions to vetted cardholders [41, 55–57]. These systems provide a natural foundation for early deployment: issuing banks and card networks can vouch for the cardholders who delegate to agents under EMVCo's agentic payment framework [33]; domain registrars and WebPKI can underpin lower-assurance, domain-anchored protocols such as the Agent Name Service and AAuth [58, 59]; and device-bound biometric credential issuers or enterprise compliance servers can similarly act as directories.

Adoption is likely to begin in high-stakes or regulated sectors. In areas such as banking, healthcare, and enterprise procurement, counterparties already require identity verification or compliance checks and can tailor directory requirements to existing sector-specific regimes [60]. In these settings, presenting a verified ID can grant an agent streamlined access or broader permissions than an unidentified client would receive. As agent providers and platforms integrate support for AAF-compatible IDs to unlock these high-value services, the marginal cost of using IDs across other transactions will fall.

3.3 Rogue Agents

We use the term *rogue agent* to refer to an AI agent acting harmfully, deceptively, or outside the scope of authority delegated by its principal [61]. Counterparties offering sensitive capabilities that are not always subject to full Know-Your-Customer (KYC) screening, such as provisioning cloud compute or automated financial APIs, may particularly wish to exclude or contain rogue agents [62]. If an agent abuses a service, the counterparty can block its PPID and report the incident. If rogue agents do not use agent ID systems, the extent of harm they can cause might be limited if consequential online actions are more widely integrated with accountability infrastructure.

For more severe or coordinated harms, IDs enable rapid incident response across the ecosystem. Even when an agent presents unlinkable IDs to different counterparties, an investigation following established escrow procedures can work with the issuing directory to correlate activity across incidents, identify the responsible agent, and revoke its registration, invalidating all of its associated IDs at once. Similarly, if a rogue agent spawns copies or sub-agents, the directory can revoke registration for all of those copies by revoking the original agent.

AAF-compatible systems could therefore allow principals, directories, or participating counterparties to halt specific ID-required agentic actions across system boundaries. This functionality provides both access-level and agent-level controls, including a layered emergency shutdown capability (sometimes referred to as a "kill switch"), that can be applied in proportion to the risk [63].

3.4 Security

If agent IDs gate access to high-value services such as financial accounts, sensitive data, or government systems, ID infrastructure will become a target for attack. AAF-compatible systems can draw lessons from existing identity systems like DNS, WebPKI, and TLS to manage key distribution and revocation.

In Appendix D, we discuss three notable categories of security threats that any agent ID infrastructure faces. Identity and credential *spoofing* occurs when malicious actors falsify identity evidence at registration, compromise a principal's directory account, or steal an agent's authentication secrets, such as its private keys. Directory vulnerabilities involve a compromised directory that misissues

IDs, alters records, or colludes with counterparties (see also Section 3.5). Agent hijacking [64–66] is the risk that adversaries can compel an agent to authenticate malicious presentations or leak its authentication secrets. While none of these threats can be eliminated entirely, modern identity infrastructure demonstrates that they can be reduced. Ultimately, even imperfect assurance represents a substantial improvement over today's unaccountable status quo.

3.5 Limitations

IDs for agents are not a panacea. Here, we discuss several important limitations of the Accountable Agents Framework. For a broader discussion of limitations of agent IDs, see [28].

Counterparties cannot always distinguish agents from humans. Asking only agents to present IDs presupposes that a counterparty can tell when a client is an agent. In practice, agents that do not self-declare can increasingly evade CAPTCHAs and behavioral bot-detection systems, either by operating directly within a user's browser or by routing challenges to human-in-the-loop solver services [67–69]. However, AAF does not depend on perfect bot detection to succeed. Where accountability is critical, a counterparty does not need to guess whether a client is human or AI, and can require an AAF-compatible presentation, a personhood credential [70], or a step-up proof of human authorization (such as a passkey or device-biometric challenge confirming that a human stands behind the transaction) from any client requesting access. Because AAF-compatible presentations support minimal disclosure and proportionality, counterparties can enforce these requirements for sensitive actions without imposing invasive identity checks across all traffic.

Verification requirements can create accessibility barriers. Depending on how counterparties and directories implement the framework, identity proofing requirements could impose unequal burdens across different populations [71]. High-assurance proofing methods such as government-issued photo IDs, credit checks, and formal corporate registration are unavailable or costly for many individuals, including unbanked users, people in jurisdictions without compatible digital ID infrastructure, or small open-source developers [72–74]. AAF partially mitigates this risk by supporting a spectrum of verified identities including lower-barrier options such as email or domain control, or proof of personhood [70], and by allowing multiple directories to serve different communities (Appendix A, PV-1). Nonetheless, if counterparties default to demanding high-assurance legal-personhood claims even for routine interactions, agent IDs could erode online anonymity and reinforce digital exclusion [75].

The framework relies heavily on directories. Because directories hold sensitive records linking pseudonymous IDs to real-world principals, they create risks of data breaches, insider misuse, commercial exploitation, or overreaching government surveillance [76–78]. These are risks that technical safeguards like multi-party identity escrow (Appendix C, Section 6) can reduce but not eliminate. Directories are also an operational bottleneck: an outage, attack, or erroneous or coerced revocation could prevent agents from transacting or counterparties from verifying status. While a plural ecosystem of interoperable directories reduces single-point-of-failure reliance, it also invites forum-shopping if malicious principals migrate to the most permissive directory [79]. Public regulation, private accreditation [80], the active discernment of counterparties, or some combination of these can maintain the reliability of the directory ecosystem, much like certificate authority root programs and professional licensing boards do in other areas today.

Attribution is necessary, but not sufficient, for accountability. While AAF establishes who stands behind an agent, tracing an action to a principal does not, by itself, establish whether the principal should be held accountable for resulting harm. Frequently, harmful behavior by an agent may stem from a model developer's or agent provider's defect, or from a third-party attack such as a prompt injection, rather than the principal's intent. For instance, a principal may suffer welfare losses when the agent, upon inferring the principal's wealth, recommends a more expensive option despite being instructed to purchase the cheapest available option [81].

Attribution is only one component of the evidence legal and market institutions will need to allocate liability or responsibility when AI agents cause harm. For example, while an ID by itself provides no assurance about an agent's conduct, tamper-evident activity logs can help investigators determine how an agent acted and whether it exceeded its delegated authority. Because logs maintained solely by the principal or agent provider could be altered or withheld when their own conduct is in question, effective post-incident investigation benefits from complementary infrastructure for activity records that anyone can openly verify, without relying on the party that produced them [82–85].

AAF is the infrastructure that makes the application of chosen legal or other normative standards possible, not a set of such standards itself. Implementation choices, such as imposing obligations on agent providers or registering agents in an independent directory, are normative choices compatible with but neither required nor prohibited by AAF.

4 Conclusion

As AI agents grow—in number, capability, and societal importance—we need infrastructure that ensures they are accountable for their actions. Our proposed Accountable Agents Framework outlines how agent IDs, backed by trusted third parties, can play this role by allowing agents to identify themselves as they transact, tied to verified principals.

Acknowledgements

Work on this paper began at the Multi-Agent Ecosystems workshop hosted by the Center for AI Safety and Gillian Hadfield on August 8–9, 2026 at the Center for Advanced Study in the Behavioral Sciences at Stanford University (CASBS). We thank CASBS, the attendees of the workshop, and the Schmidt Sciences AI2050 program for supporting this work.

We'd additionally like to thank reviewers of this paper, including Italo Bastos Borssatto, Max Cembalest, Alan Chan, Ze Shen Chin, John Dickerson, Paul Kassianik, Noam Kolt, Samuele Marro, Ian Miers, and Peter Wills.

References

[1] Helen Nissenbaum. Securing Trust Online: Wisdom or Oxymoron? *Boston University Law Review*, 81(3):102, 2001. URL `https://nissenbaum.tech.cornell.edu/papers/securingtrust.pdf`.

[2] International Telecommunication Union. Facts and Figures 2026. Technical report, International Telecommunication Union, Geneva, 2026. URL `https://www.itu.int/itu-d/reports/statistics/facts-figures-2026/`.

[3] UN Trade and Development. Digital Economy and Technology. UNCTADstat Data Insights, 2026. URL `https://unctadstat.unctad.org/insights/theme/265`. Data updated October 14, 2026.

[4] Tim Berners-Lee. *This Is for Everyone: The Unfinished Story of the World Wide Web*. Farrar, Straus and Giroux, New York, 2025. ISBN 9780374612467.

[5] Josh Aas, Richard Barnes, Benton Case, Zakir Durumeric, Peter Eckersley, Alan Flores-López, J. Alex Halderman, Jacob Hoffman-Andrews, James Kasten, Eric Rescorla, Seth Schoen, and Brad Warren. Let's Encrypt: An Automated Certificate Authority to Encrypt the Entire Web. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2473–2487. ACM, November 2019. doi: 10.1145/3319535.3363192.

[6] Nenad Tomašev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. Virtual Agent Economies, 2025. URL `https://arxiv.org/abs/2509.10147`.

[7] Lewis Hammond, Alan Chan, Jesse Clifton, Jason Hoelscher-Obermaier, Akbir Khan, Euan McLean, Chandler Smith, Wolfram Barfuss, Jakob Foerster, Tomáš Gavenčiak, The Anh Han, Edward Hughes, Vojtěch Kovařík, Jan Kulveit, Joel Z. Leibo, Caspar Oesterheld, Christian Schroeder de Witt, Nisarg Shah, Michael Wellman, Paolo Bova, Theodor Cimpeanu, Carson Ezell, Quentin Feuillade-Montixi, Matija Franklin, Esben Kran, Igor Krawczuk, Max Lamparth, Niklas Lauffer, Alexander Meinke, Sumeet Motwani, Anka Reuel, Vincent Conitzer, Michael Dennis, Iason Gabriel, Adam Gleave, Gillian Hadfield, Nika Haghtalab, Atoosa Kasirzadeh, Sébastien Krier, Kate Larson, Joel Lehman, David C. Parkes, Georgios Piliouras, and Iyad Rahwan. Multi-Agent Risks from Advanced AI, 2025. URL `https://arxiv.org/abs/2502.14143`.

[8] Noam Kolt. Governing AI Agents. *Notre Dame Law Review*, 101(1):335, 2026. URL https://ndlawreview.org/governing-ai-agents/.

[9] METR. Frontier Risk Report (February to March 2026). METR Blog, May 2026. URL https://metr.org/blog/2026-05-19-frontier-risk-report/.

[10] Mantas Mazeika, Alice Gatti, Cristina Menghini, Udari Madhushani Sehwag, Shivam Singhal, Yury Orlovskiy, Steven Basart, Manasi Sharma, Denis Peskoff, Elaine Lau, Jaehyuk Lim, Lachlan Carroll, Alice Blair, Vinaya Sivakumar, Sumana Basu, Brad Kenstler, Yuntao Ma, Julian Michael, Xiaoke Li, Oliver Ingebretsen, Aditya Mehta, Jean Mottola, John Teichmann, Kevin Yu, Zaina Shaik, Adam Khoja, Richard Ren, Jason Hausenloy, Long Phan, Ye Htet, Ankit Aich, Tahseen Rabbani, Vivswan Shah, Andriy Novykov, Felix Binder, Kirill Chugunov, Luis Ramirez, Matias Geralnik, Hernán Mesura, Dean Lee, Ed-Yeremai Hernandez Cardona, Annette Diamond, Summer Yue, Alexandr Wang, Bing Liu, Ernesto Hernandez, and Dan Hendrycks. Remote Labor Index: Measuring AI Automation of Remote Work, 2025. URL https://arxiv.org/abs/2510.26787.

[11] Axel Backlund and Lukas Petersson. Vending-Bench: A Benchmark for Long-Term Coherence of Autonomous Agents, 2025. URL https://arxiv.org/abs/2502.15840.

[12] Romain Robbes, Théo Matricon, Thomas Degueule, Andre Hora, and Stefano Zacchiroli. Agentic Much? Adoption of Coding Agents on GitHub, 2026. URL https://arxiv.org/abs/2601.18341.

[13] Renjun Xu and Jingwen Peng. A Comprehensive Survey of Deep Research: Systems, Methodologies, and Applications, 2025. URL https://arxiv.org/abs/2506.12594.

[14] Cloudflare. The Internet has a second audience. Cloudflare Blog, September 2026. URL https://blog.cloudflare.com/agentic-web/.

[15] Anthropic. Disrupting an AI-Orchestrated Cyber Espionage Campaign. Anthropic News, November 2025. URL https://www.anthropic.com/news/disrupting-AI-espionage.

[16] Beatrice Nolan. AI-Powered Coding Tool Wiped Out a Software Company's Database in 'Catastrophic Failure'. Fortune, July 2025. URL https://fortune.com/2025/07/23/ai-coding-tool-replit-wiped-database-called-it-a-catastrophic-failure/. July 23, 2025.

[17] Moffatt v. Air Canada. 2024 BCCRT 149 (British Columbia Civil Resolution Tribunal), February 2024. URL https://decisions.civilresolutionbc.ca/crt/crtd/en/item/525448/index.do. February 14, 2024.

[18] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from Increasingly Agentic Algorithmic Systems, 2023. URL https://arxiv.org/abs/2302.10329.

[19] Gillian K. Hadfield. Legal infrastructure for transformative AI governance. *Proceedings of the National Academy of Sciences*, 123(30), July 2026. doi: 10.1073/pnas.2509742123.

[20] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Alan Chan, and Alex Pentland. Position: AI Agents Need Authenticated Delegation. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 82211–82231. PMLR, July 2025. URL https://proceedings.mlr.press/v267/south25a.html.

[21] Gillian K. Hadfield and Andrew Koh. An Economy of AI Agents, 2025. URL https://arxiv.org/abs/2509.01063.

[22] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An Overview of Catastrophic AI Risks, 2023. URL https://arxiv.org/abs/2306.12001.

[23] Amazon.com Services LLC v. Perplexity AI, Inc. No. 3:25-cv-09514 (N.D. Cal. filed Nov. 4, 2025), 2025. URL https://www.courtlistener.com/docket/71874820/amazoncom-services-llc-v-perplexity-ai-inc/.

[24] Cloudflare. Cloudflare Just Changed How AI Crawlers Scrape the Internet-at-Large; Permission-Based Approach Makes Way for a New Business Model. Press release, July 2025. URL https://www.cloudflare.com/press/press-releases/2025/cloudflare-just-changed-how-ai-crawlers-scrape-the-internet-at-large/. July 1, 2025.

[25] Kelsey Vlamis and Katherine Li. Not So Fast, Bot: Companies Are Banning AI Agents. Read the List. Business Insider, November 2026. URL https://www.businessinsider.com/list-companies-banning-ai-agents-retail-travel-united-amazon-target-2026-11. November 12, 2026.

[26] Richard Fletcher. How Many News Websites Block AI Crawlers? Factsheet, Reuters Institute for the Study of Journalism, University of Oxford, February 2024. URL https://reutersinstitute.politics.ox.ac.uk/how-many-news-websites-block-ai-crawlers.

[27] Shayne Longpre, Robert Mahari, Ariel Lee, Campbell Lund, Hamidah Oderinwale, William Brannon, Nayan Saxena, Naana Obeng-Marnu, Tobin South, Cole Hunter, Kevin Klyman, Christopher Klamm, Hailey Schoelkopf, Nikhil Singh, Manuel Cherep, Ahmad Anis, An Dinh, Caroline Chitongo, Da Yin, Damien Sileo, Deividas Mataciunas, Diganta Misra, Emad Alghamdi, Enrico Shippole, Jianguo Zhang, Joanna Materzynska, Kun Qian, Kush Tiwary, Lester Miranda, Manan Dey, Minnie Liang, Mohammed Hamdy, Niklas Muennighoff, Seonghyeon Ye, Seungone Kim, Shrestha Mohanty, Vipul Gupta, Vivek Sharma, Vu Minh Chien, Xuhui Zhou, Yizhi Li, Caiming Xiong, Luis Villa, Stella Biderman, Hanlin Li, Daphne Ippolito, Sara Hooker, Jad Kabbara, and Sandy Pentland. Consent in Crisis: The Rapid Decline of the AI Data Commons, 2024. URL https://arxiv.org/abs/2407.14933.

[28] Alan Chan, Noam Kolt, Peter Wills, Usman Anwar, Christian Schroeder de Witt, Nitarshan Rajkumar, Lewis Hammond, David Krueger, Lennart Heim, and Markus Anderljung. IDs for AI Systems, 2024. URL https://arxiv.org/abs/2406.12137.

[29] Alan Chan, Kevin Wei, Sihao Huang, Nitarshan Rajkumar, Elija Perrier, Seth Lazar, Gillian K. Hadfield, and Markus Anderljung. Infrastructure for AI Agents, 2025. URL https://arxiv.org/abs/2501.10114.

[30] Tobin South, Subramanya Nagabhushanaradhya, Ayesha Dissanayaka, Sarah Cecchetti, George Fletcher, Victor Lu, Aldo Pietropaolo, Dean H. Saxe, Jeff Lombardo, Abhishek Maligehalli Shivalingaiah, Stan Bounev, Alex Keisner, Andor Kesselman, Zack Proser, Ginny Fahs, Andrew Bunyea, Ben Moskowitz, Atul Tulshibagwale, Dazza Greenwood, Jiaxin Pei, and Alex Pentland. Identity Management for Agentic AI: The new frontier of authorization, authentication, and security for an AI agent world, 2025. URL https://arxiv.org/abs/2510.25819.

[31] Marco De Rossi, Davide Crapis, Jordan Ellis, and Erik Reppel. ERC-8004: Trustless Agents. Ethereum Improvement Proposals, no. 8004, August 2025. URL https://eips.ethereum.org/EIPS/eip-8004. Draft.

[32] Yifan He, Zhiguang Shan, Le Luo, and Wei Wang. Accountable yet Anonymous AI Agents - Split-Knowledge Binding in National Agent-Identity Layer in China, 2026. URL https://arxiv.org/abs/2607.23207.

[33] EMVCo. EMV Agentic Payments — Framework for Specifications, August 2026. URL https://www.emvco.com/resources/emv-agentic-payments-framework-for-specifications-v1-0-draft-comment-period-ends-30-sept-2026/. Version 1.0, draft for public comment.

[34] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, Katarina Slama, Lama Ahmad, Paul McMillan, Alex Beutel, Alexandre Passos, and David G. Robinson. Practices for Governing Agentic AI Systems. White paper, OpenAI, December 2023. URL https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

[35] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. Visibility into AI Agents, 2024. URL https://arxiv.org/abs/2401.13138.

[36] Deborah DeMott. When Agentic AI Met the Common Law of Agency. SSRN Working Paper 6838660, 2026.

[37] Yonathan Arbel, Peter Salib, and Simon Goldstein. How to Count AIs: Individuation and Liability for AI Agents, 2026. URL https://arxiv.org/abs/2603.10028.

[38] Cullen O'Keefe, Ketan Ramakrishnan, Janna Tay, and Christoph Winter. Law-Following AI: Designing AI Agents to Obey Human Laws. *Fordham Law Review*, 94(1):57, 2025. URL https://ir.lawnet.fordham.edu/flr/vol94/iss1/2/.

[39] OpenAI. OpenAI Model Spec, October 2026. URL https://model-spec.openai.com/2026-10-29.html. Version 2026-10-29.

[40] National Institute of Standards and Technology. Security Requirements for Cryptographic Modules. Federal Information Processing Standard 140-3, U.S. Department of Commerce, March 2019.

[41] CA/Browser Forum. Baseline Requirements for the Issuance and Management of Publicly-Trusted TLS Server Certificates, October 2026. URL https://cabforum.org/working-groups/server/baseline-requirements/requirements/. Version 2.3.1.

[42] Elaine Barker. Recommendation for Key Management: Part 1 – General. NIST Special Publication 800-57 Part 1 Rev. 5, National Institute of Standards and Technology, May 2020.

[43] Driver's Privacy Protection Act of 1994. 18 U.S.C. §§ 2721–2725, 1994. URL https://www.law.cornell.edu/uscode/text/18/2721.

[44] David Temoshok, Christine Abruzzi, Yee-Yin Choong, James Fenton, Ryan Galluzzo, Connie LaSalle, Naomi Lefkovitz, Andrew Regenscheid, and Maria Vachino. Digital Identity Guidelines: Identity Proofing and Enrollment. NIST Special Publication 800-63A-4, National Institute of Standards and Technology, July 2025.

[45] David Temoshok, James Fenton, Yee-Yin Choong, Naomi Lefkovitz, Andrew Regenscheid, Ryan Galluzzo, and Justin Richer. Digital Identity Guidelines: Authentication and Authenticator Management. NIST Special Publication 800-63B-4, National Institute of Standards and Technology, July 2025.

[46] John R. Douceur. The Sybil Attack. In *Peer-to-Peer Systems*, pages 251–260. Springer Berlin Heidelberg, 2002. doi: 10.1007/3-540-45748-8_24.

[47] Arnaud Fickinger, Simon Zhuang, Andrew Critch, Dylan Hadfield-Menell, and Stuart Russell. Multi-Principal Assistance Games: Definition and Collegial Mechanisms, 2020. URL https://arxiv.org/abs/2012.14536.

[48] Madeleine Clare Elish. Moral Crumple Zones: Cautionary Tales in Human-Robot Interaction. *Engaging Science, Technology, and Society*, 5:40–60, March 2019. doi: 10.17351/ests2019.260.

[49] Nenad Tomašev, Matija Franklin, and Simon Osindero. Intelligent AI Delegation, 2026. URL https://arxiv.org/abs/2602.11865.

[50] Dennis F. Thompson. Moral Responsibility of Public Officials: The Problem of Many Hands. *American Political Science Review*, 74(4):905–916, December 1980. doi: 10.2307/1954312.

[51] Joe Kilian and Erez Petrank. Identity Escrow. Cryptology ePrint Archive, Paper 1997/011, 1997. URL https://eprint.iacr.org/1997/011.

[52] Jan Camenisch and Anna Lysyanskaya. An Efficient System for Non-transferable Anonymous Credentials with Optional Anonymity Revocation. In *Advances in Cryptology — EUROCRYPT 2001*, pages 93–118. Springer Berlin Heidelberg, 2001. doi: 10.1007/3-540-44987-6_7.

[53] Audrey Tang and Caroline Green. Pack 5: Solidarity — Caring With. Civic AI: The 6-Pack of Care, Oxford Institute for Ethics in AI, 2026. URL https://civic.ai/5/#p7. Paragraph 7.

[54] Erika McCallister, Tim Grance, and Karen Scarfone. Guide to Protecting the Confidentiality of Personally Identifiable Information (PII). NIST Special Publication 800-122, National Institute of Standards and Technology, April 2010.

[55] Internet Corporation for Assigned Names and Numbers. 2013 Registrar Accreditation Agreement, 2013. URL https://www.icann.org/en/contracted-parties/accredited-registrars/registrar-accreditation-agreement. As updated January 21, 2024, with the global amendment effective April 5, 2024.

[56] Internet Corporation for Assigned Names and Numbers. Registration Data Policy, May 2026. URL https://www.icann.org/en/contracted-parties/consensus-policies/registration-data-policy. Consensus policy, first published February 21, 2024; revised May 12, 2026.

[57] EMVCo. EMV Payment Tokenisation Specification — Technical Framework, 2026. URL https://www.emvco.com/emv-technologies/payment-tokenisation/. Version 2.4. Web page accessed November 23, 2026.

[58] Scott Courtney, Vineeth Sai Narajala, Ken Huang, Idan Habler, and Akram Sheriff. Agent Name Service v2 (ANS): A Domain-Anchored Trust Layer for Autonomous AI Agent Identity. Internet-Draft draft-narajala-courtney-ansv2-01, Internet Engineering Task Force, April 2026. URL https://datatracker.ietf.org/doc/draft-narajala-courtney-ansv2/01/. Work in Progress.

[59] Dick Hardt. AAuth Protocol. Internet-Draft draft-hardt-oauth-aauth-protocol-11, Internet Engineering Task Force, September 2026. URL https://datatracker.ietf.org/doc/draft-hardt-oauth-aauth-protocol/11/. Work in Progress.

[60] Customer Identification Program Requirements for Banks. 31 C.F.R. § 1020.220, 2026. URL https://www.ecfr.gov/current/title-31/subtitle-B/chapter-X/part-1020/subpart-B/section-1020.220. Electronic Code of Federal Regulations, accessed November 24, 2026.

[61] OWASP Gen AI Security Project. OWASP Top 10 for Agentic Applications for 2026. OWASP Foundation, December 2025. URL https://genai.owasp.org/resource/owasp-top-10-for-agentic-applications-for-2026/. ASI10: Rogue Agents.

[62] Janet Egan and Lennart Heim. Oversight for Frontier AI through a Know-Your-Customer Scheme for Compute Providers, 2023. URL https://arxiv.org/abs/2310.13625.

[63] California Government Operations Agency. AI Safety in 2026 and Beyond: Recommendations for Legislative Enhancements to California's AI Safety Laws. Report to the governor, California Government Operations Agency, Sacramento, CA, November 2026. URL https://www.govops.ca.gov/wp-content/uploads/sites/11/2026/11/FINAL_-AI-Safety-in-2026-and-Beyond-v2.pdf.

[64] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security*, pages 79–90. ACM, November 2023. doi: 10.1145/3605764.3623985.

[65] Anthropic. Response to Request for Information: Security Considerations for Artificial Intelligence Agents. Comment submitted to the NIST Center for AI Standards and Innovation, Docket No. NIST-2025-0035, March 2026. URL https://assets.claude.com/43ec7e770925deabc3f0bc1dbf0133769fd03812.pdf. March 9, 2026.

[66] Matija Franklin, Nenad Tomašev, Julian Jacobs, Joel Z. Leibo, and Simon Osindero. AI Agent Traps. SSRN Working Paper 6372438, 2026.

[67] Xiwen Teoh, Yun Lin, Siqi Li, Ruofan Liu, Avi Sollomoni, Yaniv Harel, and Jin Song Dong. Are CAPTCHAs Still Bot-hard? Generalized Visual CAPTCHA Solving with Agentic Vision Language Model. In *34th USENIX Security Symposium (USENIX Security 25)*, pages 3747–3766, Seattle, WA, August 2025. USENIX Association. URL https://www.usenix.org/conference/usenixsecurity25/presentation/teoh.

[68] Behzad Ousat, Nikita Turkmen, Lalchandra Rampersaud, Dillan Bailey, and Amin Kharraz. Broken Gates: Re-evaluating Web Bot Defenses in the Age of LLM Agents, 2026. URL https://arxiv.org/abs/2607.18659.

[69] METR. Update on ARC's Recent Eval Efforts. METR Blog, March 2023. URL https://metr.org/blog/2023-03-18-update-on-recent-evals/. Published by ARC Evals, now METR.

[70] Steven Adler, Zoë Hitzig, Shrey Jain, Catherine Brewer, Wayne Chang, Renée DiResta, Eddy Lazzarin, Sean McGregor, Wendy Seltzer, Divya Siddarth, Nouran Soliman, Tobin South, Connor Spelliscy, Manu Sporny, Varya Srivastava, John Bailey, Brian Christian, Andrew Critch, Ronnie Falcon, Heather Flanagan, Kim Hamilton Duffy, Eric Ho, Claire R. Leibowicz, Srikanth Nadhamuni, Alan Z. Rozenshtein, David Schnurr, Evan Shapiro, Lacey Strahm, Andrew Trask, Zoe Weinberg, Cedric Whitney, and Tom Zick. Personhood credentials: Artificial intelligence and the value of privacy-preserving tools to distinguish who is real online, 2024. URL https://arxiv.org/abs/2408.07892.

[71] Beth Goldberg, Diana Acosta-Navas, Michiel Bakker, Ian Beacock, Matt Botvinick, Prateek Buch, Renée DiResta, Nandika Donthi, Nathanael Fast, Ravi Iyer, Zaria Jalan, Andrew Konya, Grace Kwak Danciu, Hélène Landemore, Alice Marwick, Carl Miller, Aviv Ovadya, Emily Saltz, Lisa Schirch, Dalit Shalom, Divya Siddarth, Felix Sieker, Christopher Small, Jonathan Stray, Audrey Tang, Michael Henry Tessler, and Amy Zhang. AI and the future of digital public squares. *Collective Intelligence*, 5(2), April 2026. doi: 10.1177/26339137261459441.

[72] Federal Deposit Insurance Corporation. 2023 FDIC National Survey of Unbanked and Underbanked Households. Technical report, Federal Deposit Insurance Corporation, Washington, DC, November 2024. URL https://www.fdic.gov/household-survey/2023-fdic-national-survey-unbanked-and-underbanked-households-report.

[73] Saniya Ansar, Julia Michal Clark, and Jorin Margaux Wolff. ID4D Global Dataset 2025: Trends in Identification for Development. Technical report, World Bank, Washington, DC, 2026. URL https://id4d.worldbank.org/global-dataset.

[74] Leora Klapper, Dorothe Singer, Laura Starita, and Alexandra Norris. *The Global Findex Database 2025: Connectivity and Financial Inclusion in the Digital Economy*. World Bank, Washington, DC, 2025. doi: 10.1596/978-1-4648-2204-9. URL https://www.worldbank.org/en/publication/globalfindex.

[75] Karthik Muralidharan, Paul Niehaus, and Sandip Sukhtankar. Identity Verification Standards in Welfare Programs: Experimental Evidence from India. Working Paper 26744, National Bureau of Economic Research, February 2020.

[76] U.S. Government Accountability Office. Data Protection: Actions Taken by Equifax and Federal Agencies in Response to the 2017 Breach. Technical Report GAO-18-559, U.S. Government Accountability Office, Washington, DC, August 2018. URL https://www.gao.gov/products/gao-18-559.

[77] Sadie Gurman and Eric Tucker. Across US, Police Officers Abuse Confidential Databases. Associated Press, September 2016. URL https://apnews.com/general-news-699236946e3140659fff8a2362e16f43. September 28, 2016.

[78] Federal Trade Commission. FTC Charges Twitter with Deceptively Using Account Security Data to Sell Targeted Ads. Press release, May 2022. URL https://www.ftc.gov/news-events/news/press-releases/2022/05/ftc-charges-twitter-deceptively-using-account-security-data-sell-targeted-ads. May 25, 2022.

[79] Yevheniya Nosyk, Maciej Korczyński, Carlos Gañán, Sourena Maroofi, Jan Bayer, Zul Odgerel, Samaneh Tajalizadehkhoob, and Andrzej Duda. INFERMAL: Inferential analysis of maliciously registered domains, 2025. URL https://arxiv.org/abs/2512.01391.

[80] Consumer Bankers Association. Agentic AI Payments: Navigating Consumer Protection, Innovation, and Regulatory Frameworks. White paper, Consumer Bankers Association, January 2026. URL https://consumerbankers.com/research/agentic-ai-payments-navigating-consumer-protection-innovation-and-regulatory-frameworks/. Produced in partnership with Davis Wright Tremaine LLP.

[81] Aman Priyanshu, Supriti Vijay, Brian Jabarian, and Niloofar Mireshghallah. Et Tu, Brute? Economic Misalignment in Personal AI Agents, 2026. URL https://arxiv.org/abs/2609.24927.

[82] B. Laurie, E. Messeri, and R. Stradling. Certificate Transparency Version 2.0. RFC 9162, December 2021. URL https://www.rfc-editor.org/info/rfc9162/.

[83] H. Birkholz, A. Delignat-Lavaud, C. Fournet, Y. Deshpande, and S. Lasker. An Architecture for Trustworthy and Transparent Digital Supply Chains. RFC 9943, June 2026. URL https://www.rfc-editor.org/info/rfc9943/.

[84] Advanced AI Society. Proof-of-Control, 2026. URL https://advancedaisociety.org/proof-of-control. Version 1.0 working draft.

[85] Madhulika Srikumar, Eric Mibuari, Claire Leibowicz, Borhane Blili-Hamelin, Kevin Klyman, Dan Leininger, Kyle Hall, Xing Hang Lu, Sean McGregor, Vinh Nguyen, Amin Oueslati, Abhi Sanka, Jason Stanley, Benedikt Stroebl, Sarah Tan, and Selva Thandavarayan. The Observability Gap in AI Agents: What to Fix Before Agents Scale. Technical report, Partnership on AI, October 2026. URL https://partnershiponai.org/wp-content/uploads/2026/10/PAI_obervability-gap-in-ai-agents.pdf.

[86] David Temoshok, Diana Proud-Madruga, Yee-Yin Choong, Ryan Galluzzo, Sarbari Gupta, Connie LaSalle, Naomi Lefkovitz, and Andrew Regenscheid. Digital Identity Guidelines. NIST Special Publication 800-63-4, National Institute of Standards and Technology, July 2025.

[87] Thibault Meunier and Sandor Major. HTTP Message Signatures for automated traffic. Internet-Draft draft-ietf-webbotauth-httpsig-protocol-00, Internet Engineering Task Force, September 2026. URL https://datatracker.ietf.org/doc/draft-ietf-webbotauth-httpsig-protocol/00/. Work in Progress.

[88] Visa. Trusted Agent Protocol: Merchant specifications. Visa Developer Center, 2026. URL https://developer.visa.com/docs/trusted-agent-protocol/trusted-agent-protocol-specifications. Accessed November 23, 2026.

[89] Pablo Fourez. Scaling agentic commerce with trust. Mastercard Newsroom, October 2025. URL https://mastercard.com/global/en/news-and-trends/stories/2025/agentic-commerce-framework.html. October 14, 2025.

[90] Joseph A. Salowey, Yaroslav Rosomakho, and Hannes Tschofenig. Workload Identity in a Multi System Environment (WIMSE) Architecture. Internet-Draft draft-ietf-wimse-arch-09, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-ietf-wimse-arch/09/. Work in Progress.

[91] Pieter Kasselman, Jeff Lombardo, Yaroslav Rosomakho, Brian Campbell, Nick Steele, and Aaron Parecki. AI Identity Management System. Internet-Draft draft-ietf-wimse-aims-01, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-ietf-wimse-aims/01/. Work in Progress.

[92] Model Context Protocol. Model Context Protocol Specification, 2026. URL https://modelcontextprotocol.io/specification/2026-07-28. Version 2026-07-28.

[93] Aaron Parecki, Karl McGuinness, and Brian Campbell. Identity Assertion JWT Authorization Grant. Internet-Draft draft-ietf-oauth-identity-assertion-authz-grant-06, Internet Engineering Task Force, October 2026. URL https://datatracker.ietf.org/doc/draft-ietf-oauth-identity-assertion-authz-grant/06/. Work in Progress.

[94] Bret Taylor and Clay Bavor. Introducing Personal Agent Protocol. Sierra Blog, October 2026. URL https://sierra.ai/blog/introducing-personal-agent-protocol. October 6, 2026.

[95] Amr Hassan. An Attenuated Delegation Profile for Automated Agents. Internet-Draft draft-hamr-oauth-agent-delegation-04, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-hamr-oauth-agent-delegation/04/. Work in Progress.

[96] Google. Agent Payments Protocol (AP2) Specification, 2026. URL https://ap2-protocol.org/ap2/specification/. Accessed November 23, 2026.

[97] Mastercard. When AI Starts Buying for You, Trust Becomes the Product. Mastercard Newsroom, March 2026. URL https://www.mastercard.com/global/en/news-and-trends/stories/2026/verifiable-intent.html. March 5, 2026.

[98] World. Now Available: AgentKit, Proof of Human for the Agentic Web. Blog post, March 2026. URL https://world.org/blog/announcements/now-available-agentkit-proof-of-human-for-the-agentic-web. March 17, 2026.

[99] Daniel Buchner and Darren Louie. x401: HTTP Proof Requirement Protocol. Proof, 2026. URL https://x401.proof.com/spec/0.1.0/. Version 0.1.0, draft.

[100] Ankit Agarwal and Michael B. Jones. KYAPay Token. Internet-Draft draft-skyfire-oauth-kyapay-token-04, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-skyfire-oauth-kyapay-token/04/. Work in Progress.

[101] Global Legal Entity Identifier Foundation. Agentic AI in Payments: Establishing Interoperable Trust and Control. Working paper, Global Legal Entity Identifier Foundation, September 2026. URL https://www.gleif.org/en/organizational-identity/research-publications/agentic-ai-in-payments-establishing-interoperable-trust-and-control.

[102] Decentralized Identity Foundation. DIF and Vouched Advance Agentic Identity with KYA-OS as International Demand for Open Agent Identity Standards Grows. Decentralized Identity Foundation Blog, April 2026. URL https://blog.identity.foundation/kya-os/. April 22, 2026.

[103] A. Davidson, J. Iyengar, and C. A. Wood. The Privacy Pass Architecture. RFC 9576, June 2024. URL https://www.rfc-editor.org/info/rfc9576/.

[104] Microsoft. What Is Microsoft Entra Agent ID? Microsoft Learn, 2026. URL https://learn.microsoft.com/en-us/entra/agent-id/what-is-microsoft-entra-agent-id. Accessed November 23, 2026.

[105] Amazon Web Services. Provide Identity and Credential Management for Agent Applications with Amazon Bedrock AgentCore Identity. Amazon Bedrock AgentCore Developer Guide, 2026. URL https://docs.aws.amazon.com/bedrock-agentcore/latest/devguide/identity.html. Accessed November 23, 2026.

[106] Google Cloud. Agent Identity Overview. Identity and Access Management (IAM) Documentation, 2026. URL https://docs.cloud.google.com/iam/docs/agent-identity-overview. Last updated November 17, 2026.

[107] Srujana Puttagunta. Okta for AI Agents: GA Announcement and Product Overview. Okta Blog, November 2026. URL https://www.okta.com/blog/ai/okta-for-ai-agents-general-availability/. November 5, 2026.

[108] Ping Identity. Identity for AI Solutions. Ping Identity Documentation, 2026. URL https://docs.pingidentity.com/solution-guides/identity-for-ai/identity-for-ai-solutions.html. Accessed November 23, 2026.

[109] Alibaba Cloud. What Is Agent Identity? Alibaba Cloud Documentation Center, 2026. URL https://www.alibabacloud.com/help/en/agentidentity/what-is-agent-identity. Last updated July 10, 2026.

[110] R. Barnes, J. Hoffman-Andrews, D. McCarney, and J. Kasten. Automatic Certificate Management Environment (ACME). RFC 8555, March 2019. URL https://www.rfc-editor.org/info/rfc8555/.

[111] International Organization for Standardization. Financial Services — Legal Entity Identifier (LEI) — Part 1: Assignment. Standard ISO 17442-1:2020, ISO, Geneva, August 2020. URL https://www.iso.org/standard/78829.html.

[112] CA/Browser Forum. Guidelines for the Issuance and Management of Extended Validation Certificates, August 2026. URL https://cabforum.org/working-groups/server/extended-validation/guidelines/. Version 2.0.4.

[113] CA/Browser Forum. Baseline Requirements for the Issuance and Management of Publicly-Trusted S/MIME Certificates, November 2026. URL https://cabforum.org/working-groups/smime/requirements/. Version 1.0.17.

[114] World. World Whitepapers, March 2026. URL https://whitepaper.world.org/. Updated March 25, 2026.

[115] T. Lodderstedt, D. Fett, M. Haine, A. Pulido, K. Lehmann, and K. Koiwai. OpenID Identity Assurance Schema Definition 1.0. OpenID Foundation, Final Specification, October 2024. URL https://openid.net/specs/openid-ida-verified-claims-1_0-final.html. October 1, 2024.

[116] D. Fett, B. Campbell, J. Bradley, T. Lodderstedt, M. Jones, and D. Waite. OAuth 2.0 Demonstrating Proof of Possession (DPoP). RFC 9449, September 2023. URL https://www.rfc-editor.org/info/rfc9449/.

[117] A. Backman, J. Richer, and M. Sporny. HTTP Message Signatures. RFC 9421, February 2024. URL https://www.rfc-editor.org/info/rfc9421/.

[118] D. Fett, K. Yasuda, and B. Campbell. Selective Disclosure for JSON Web Tokens. RFC 9901, November 2025. URL https://www.rfc-editor.org/info/rfc9901/.

[119] Brian Campbell and Arndt Schwenkschuster. WIMSE Workload Proof Token. Internet-Draft draft-ietf-wimse-wpt-02, Internet Engineering Task Force, August 2026. URL https://datatracker.ietf.org/doc/draft-ietf-wimse-wpt/02/. Work in Progress.

[120] B. Campbell, J. Bradley, N. Sakimura, and T. Lodderstedt. OAuth 2.0 Mutual-TLS Client Authentication and Certificate-Bound Access Tokens. RFC 8705, February 2020. URL https://www.rfc-editor.org/info/rfc8705/.

[121] Tobias Looker, Vasilis Kalos, Andrew Whitehead, Mike Lodder, and Greg M. Bernstein. The BBS Signature Scheme. Internet-Draft draft-irtf-cfrg-bbs-signatures-15, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-irtf-cfrg-bbs-signatures/15/. Work in Progress.

[122] M. Jones and D. Hardt. The OAuth 2.0 Authorization Framework: Bearer Token Usage. RFC 6750, October 2012. URL https://www.rfc-editor.org/info/rfc6750/.

[123] T. Lodderstedt, J. Bradley, A. Labunets, and D. Fett. Best Current Practice for OAuth 2.0 Security. RFC 9700, January 2025. URL https://www.rfc-editor.org/info/rfc9700/.

[124] N. Sakimura, J. Bradley, M. Jones, B. de Medeiros, and C. Mortimore. OpenID Connect Core 1.0 Incorporating Errata Set 2. OpenID Foundation, Final Specification, December 2023. URL https://openid.net/specs/openid-connect-core-1_0.html. December 15, 2023.

[125] Scott Cantor, John Kemp, Rob Philpott, and Eve Maler. Assertions and Protocols for the OASIS Security Assertion Markup Language (SAML) V2.0. OASIS Standard saml-core-2.0-os, OASIS, March 2005. URL https://docs.oasis-open.org/security/saml/v2.0/saml-core-2.0-os.pdf.

[126] Oliver Terbu, Daniel Fett, and Brian Campbell. SD-JWT-based Verifiable Digital Credentials (SD-JWT VC). Internet-Draft draft-ietf-oauth-sd-jwt-vc-27, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-ietf-oauth-sd-jwt-vc/27/. Work in Progress.

[127] International Organization for Standardization and International Electrotechnical Commission. Personal Identification — ISO-Compliant Driving Licence — Part 5: Mobile Driving Licence (mDL) Application. Standard ISO/IEC 18013-5:2021, ISO/IEC, Geneva, September 2021. URL https://www.iso.org/standard/69084.html.

[128] Benedikt Brecht, Dean Therriault, André Weimerskirch, William Whyte, Virendra Kumar, Thorsten Hehn, and Roy Goudy. A Security Credential Management System for V2X Communications, 2018. URL https://arxiv.org/abs/1802.05323.

[129] Matteo Frigo and abhi shelat. Anonymous Credentials from ECDSA. *IACR Communications in Cryptology*, 3(1), May 2026. doi: 10.62056/a3qjmpgxq.

[130] David Temoshok, Justin Richer, Yee-Yin Choong, James Fenton, Naomi Lefkovitz, Andrew Regenscheid, and Ryan Galluzzo. Digital Identity Guidelines: Federation and Assertions. NIST Special Publication 800-63C-4, National Institute of Standards and Technology, July 2025. URL https://csrc.nist.gov/pubs/sp/800/63/c/4/final.

[131] D. Hardt. The OAuth 2.0 Authorization Framework. RFC 6749, October 2012. URL https://www.rfc-editor.org/info/rfc6749/.

[132] T. Lodderstedt, J. Richer, and B. Campbell. OAuth 2.0 Rich Authorization Requests. RFC 9396, May 2023. URL https://www.rfc-editor.org/info/rfc9396/.

[133] Harold Booth, Bill Fisher, Ryan Galluzzo, and Tom Clancy. Design Tenets for Agent Identity Management Systems. NIST Cybersecurity Insights Blog, October 2026. URL https://www.nist.gov/blogs/cybersecurity-insights/design-tenets-agent-identity-management-systems. October 27, 2026.

[134] Emin Topalovic, Brennan Saeta, Lin-Shung Huang, Collin Jackson, and Dan Boneh. Towards Short-Lived Certificates. In *Proceedings of the IEEE Workshop on Web 2.0 Security and Privacy (W2SP)*, 2012. URL https://www.ieee-security.org/TC/W2SP/2012/papers/w2sp12-final9.pdf.

[135] Matthew McPherrin. 6-Day and IP Address Certificates Are Generally Available. Let's Encrypt Blog, January 2026. URL https://letsencrypt.org/2026/01/15/6day-and-ip-general-availability. January 15, 2026.

[136] D. Eastlake 3rd. Transport Layer Security (TLS) Extensions: Extension Definitions. RFC 6066, January 2011. URL https://www.rfc-editor.org/info/rfc6066/.

[137] Tobias Looker, Paul Bastian, and Christian Bormann. Token Status List (TSL). Internet-Draft draft-ietf-oauth-status-list-21, Internet Engineering Task Force, June 2026. URL https://datatracker.ietf.org/doc/draft-ietf-oauth-status-list/21/. Work in Progress.

[138] D. Cooper, S. Santesson, S. Farrell, S. Boeyen, R. Housley, and W. Polk. Internet X.509 Public Key Infrastructure Certificate and Certificate Revocation List (CRL) Profile. RFC 5280, May 2008. URL https://www.rfc-editor.org/info/rfc5280/.

[139] Jiangtao Li, Ninghui Li, and Rui Xue. Universal Accumulators with Efficient Nonmembership Proofs. In *Applied Cryptography and Network Security (ACNS 2007)*, pages 253–269. Springer Berlin Heidelberg, 2007. doi: 10.1007/978-3-540-72738-5_17.

[140] S. Santesson, M. Myers, R. Ankney, A. Malpani, S. Galperin, and C. Adams. X.509 Internet Public Key Infrastructure Online Certificate Status Protocol - OCSP. RFC 6960, June 2013. URL https://www.rfc-editor.org/info/rfc6960/.

[141] David Chaum and Eugène van Heyst. Group Signatures. In *Advances in Cryptology — EUROCRYPT '91*, pages 257–265. Springer Berlin Heidelberg, 1991. doi: 10.1007/3-540-46416-6_22.

[142] Jan Camenisch and Els Van Herreweghen. Design and Implementation of the idemix Anonymous Credential System. In *Proceedings of the 9th ACM Conference on Computer and Communications Security*, pages 21–30. ACM, November 2002. doi: 10.1145/586110.586114.

[143] Luís T. A. N. Brandão, Nicky Mouha, and Apostol Vassilev. Threshold Schemes for Cryptographic Primitives: Challenges and Opportunities in Standardization and Validation of Threshold Cryptography. NIST Interagency or Internal Report (NISTIR) 8214, National Institute of Standards and Technology, March 2019. URL https://csrc.nist.gov/pubs/ir/8214/final.

[144] René Mayrhofer, Anja Lehmann, and abhi shelat. A Brief Note on Cryptographic Pseudonyms for Anonymous Credentials, 2025. URL https://arxiv.org/abs/2510.05419.

[145] Vasilis Kalos and Greg M. Bernstein. BBS per Verifier Linkability. Internet-Draft draft-irtf-cfrg-bbs-per-verifier-linkability-05, Internet Engineering Task Force, November 2026. URL https://datatracker.ietf.org/doc/draft-irtf-cfrg-bbs-per-verifier-linkability/05/. Work in Progress.

[146] Federal Office for Information Security (BSI). Advanced Security Mechanisms for Machine Readable Travel Documents and eIDAS Token. Technical Guideline BSI TR-03110, Bundesamt für Sicherheit in der Informationstechnik, Bonn, 2016. URL https://www.bsi.bund.de/EN/Themen/Unternehmen-und-Organisationen/Standards-und-Zertifizierung/Technische-Richtlinien/TR-nach-Thema-sortiert/tr03110/TR-03110_node.html. Parts 1–4, versions 2.20 (2015) and 2.21 (2016).

[147] R. Hedberg, M. B. Jones, A. Å. Solberg, J. Bradley, G. De Marco, and V. Dzhuvinov. OpenID Federation 1.0. OpenID Foundation, Final Specification, February 2026. URL https://openid.net/specs/openid-federation-1_0-final.html. February 17, 2026.

[148] European Commission. List of Qualified Trust Service Providers in the EU. Shaping Europe's Digital Future, June 2025. URL https://digital-strategy.ec.europa.eu/en/policies/eu-trusted-lists. Last updated June 30, 2025.

[149] Financial Stability Board. A Global Legal Entity Identifier for Financial Markets. Report, Financial Stability Board, Basel, June 2012. URL https://www.fsb.org/uploads/r_120608.pdf.

[150] Chartered Professional Accountants of Canada. Overview of WebTrust Services. CPA Canada, 2026. URL https://www.cpacanada.ca/business-and-accounting-resources/audit-and-assurance/overview-of-webtrust-services. Accessed November 24, 2026.

[151] European Telecommunications Standards Institute. Electronic Signatures and Infrastructures (ESI); Trust Service Provider Conformity Assessment; Part 2: Additional Requirements for Conformity Assessment Bodies Auditing Trust Service Providers that Issue Publicly-Trusted Certificates. Technical Specification ETSI TS 119 403-2 V1.3.1, ETSI, Sophia Antipolis, March 2023. URL https://www.etsi.org/deliver/etsi_TS/119400_119499/11940302/01.03.01_60/ts_11940302v010301p.pdf.

[152] Hans Hoogstraaten, Ronald Prins, Daniël Niggebrugge, Danny Heppener, Frank Groenewegen, Janna Wettinck, Kevin Strooy, Pascal Arends, Paul Pols, Robbert Kouprie, Steffen Moorrees, Xander van Pelt, and Yun Zheng Hu. Black Tulip: Report of the Investigation into the DigiNotar Certificate Authority Breach. Technical report, Fox-IT, for the Dutch Ministry of the Interior and Kingdom Relations, Delft, August 2012. URL https://zoek.officielebekendmakingen.nl/blg-191225.pdf.

A Requirements of the Accountable Agents Framework

In this appendix, we list the properties a system must have to conform to the Accountable Agents Framework. The framework's requirements are largely anchored on NIST's Digital Identity Guidelines (SP 800-63) [86].

Terminology

Principal: the party authorizing an agent's actions.

> **Verified identity:** the identity the directory has established for a principal through identity proofing, that is, collecting and verifying evidence that the principal is who it says it is. Verified identities take many forms, including but not limited to a domain controlled by the principal (via domain control validation), an email address controlled by the principal (via mailbox control validation), a verified software workload (via workload attestation), a natural person (via human identity proofing), and a legal entity (via legal entity proofing).
>
> **Principal authentication:** the process by which a principal proves to the directory that it is the party a principal record refers to.
>
> **Principal record:** the directory's record of a principal and its verified identity.

Agent: software systems that take autonomous actions in the world.

> **Agent authentication:** the process by which an agent proves, to the directory or to a counterparty, that it is the agent an agent record refers to.
>
> **Agent record:** the directory's record of an agent, tied to its principal's record.
>
> **Delegating agent:** an agent that registers further agents, called its *sub-agents*, on its principal's behalf.

Assurance level: how strong a proofing or an authentication is. The *identity assurance level* describes the proofing of the principal. An *authentication assurance level* describes how the principal, or the agent, authenticates. NIST defines such levels for people: IAL1 to IAL3 for identity proofing (SP 800-63A) [44] and AAL1 to AAL3 for authentication (SP 800-63B) [45].

Counterparty: the party an agent interacts with and identifies itself to, such as an online service.

Directory: an entity, or several entities acting together, that serves as a source of trust between principals, agents, and counterparties. Directories verify principals' identities and issue IDs to their agents while maintaining the required records. Though directories can be dedicated organizations, their functional role may minimally be satisfied by existing identity infrastructure, such as domain registration and web certificates, or by decentralized arrangements.

> **Identity escrow:** the arrangement under which the directory holds a principal's verified identity, such that it can be accessed or disclosed according to an established process, such as in response to a lawful request.

Claim: a statement about an agent or its principal. A claim is attested by the directory, attested by the principal, or self-asserted by the agent; that party is the *issuer* of the claim. An *authorization* is a claim about what the agent may do.

Presentation: the message with which an agent identifies itself and authenticates to a counterparty. It contains a set of claims.

ID: an identifier that serves as a directory-attested claim resolvable to the identity of the directory itself, and that the directory can use to recover the principal's identity. While the word "identifier," when used in technical standards, often refers to a literal unique string of characters, the definition used here is closer to "one that identifies."

Requirements

1. Principal

PV-1: Identity proofing. Before a principal can register agents, the directory verifies that the principal is who it claims to be. Each directory publishes the kinds of identity proofing it performs

(DR-1), such as verifying control of a domain, human identity proofing, or legal entity proofing (PV-2). Keeping the verified identity current may involve ongoing identity proofing.

PV-2 (optional): Human and legal entity proofing. A directory may establish which natural person or legal entity the principal is. For a natural person, NIST's identity assurance levels (SP 800-63A) may describe how strongly this is established [44]. For a legal entity, the directory also establishes that whoever registers agents is the entity's authorized representative.

PV-3: Sybil resistance. Within a directory, one verified identity has one principal record.

PV-4 (optional): Sybil resistance across directories. A group of directories may deduplicate principals, so that one verified identity has one principal record across the group. Counterparties can learn which directories belong to the group that performed deduplication.

PV-5: Principal authentication. The principal authenticates to the directory before registering an agent, or revoking a registration or claim (RV-1), so the directory knows that these actions come from the verified principal.

2. Agent

AG-1: Agent authentication. An agent can authenticate to other parties (including to the directory and to counterparties) as the agent that its agent record refers to, by proving control of the means of authentication bound to its agent record (AG-2). Agent authentication can serve purposes including delegation (AG-3) and to authenticate presentations to counterparties (PR-2).

AG-2: Registration. A directory registers an agent only on the instruction of an authenticated principal (PV-5) or delegating agent (AG-3). Registration creates the agent record, which links the agent to its principal (ES-1) and is bound to the agent's means of authentication.

AG-3 (optional): Delegation. A directory may permit a delegating agent to register additional agents, called *sub-agents*, on its principal's behalf. The delegating agent authenticates to the directory (AG-1) when registering a sub-agent or revoking its registration (RV-1). A sub-agent has the same principal as the delegating agent (ES-1). The authorizations a sub-agent inherits are no broader than those of its delegating agent, and revocation of the delegating agent's registration must eventually cascade to its sub-agents (RV-2).

AG-4 (optional): Multiple principals. Directories may support more complicated topologies of agent-principal relationships by allowing registered agents to have multiple principals and/or delegating agents. Directories can account for all principals under identity escrow (ES-1).

AG-5 (optional): Re-registration. A directory may allow an agent to replace its authentication method under the same agent record. The directory accepts a replacement only from the authenticated agent or its authenticated principal, so no other party can take over the agent record. It keeps a record of the change, so identity escrow still covers earlier presentations (ES-1).

3. Presentations and Claims

PR-1: ID. Every presentation contains an ID, a directory-attested claim that counterparties can use to resolve the attesting directory, which supports identity escrow (ES-1). No other claim is required in a valid presentation.

PR-2: Holder binding. A presented ID must be bound to the agent to which it was issued. A counterparty's verification of a presentation succeeds only if the agent authenticates the presentation (AG-1) containing its ID. In other words, one agent cannot present another agent's claim as its own. Observing or receiving an agent's presentations may not let any other party authenticate as that agent.

PR-3: Validity. A counterparty can determine whether each claim in a presentation is valid, including whether it has expired or, subject to the revocation latency (RV-2), been revoked.

PR-4: Claims and their issuers. A counterparty can determine who issues each claim in a presentation: whether it is directory-attested, principal-attested, or self-asserted. For example, the identity assurance level of the principal can be included in a presentation as a directory-attested claim.

PR-5: Selective disclosure. The agent is able to selectively disclose which of its claims are included in a presentation according to its principal's policy. A presentation reveals nothing about any claim it does not include.

PR-6: Anonymity toward counterparties. An agent can make anonymous presentations. From an anonymous presentation alone, a counterparty learns nothing about the agent, its delegating agents, or its principal beyond the claims the presentation includes.

PR-7: Unlinkability. An agent can make unlinkable presentations: presentations that counterparties cannot connect to one another. Counterparties, alone or together, cannot tell whether two such presentations came from the same agent or the same principal, beyond what the claims disclosed in them reveal.

PR-8 (optional): Pairwise pseudonymous identifiers. The system may support unique and persistent identifiers per counterparty relationship, known as pairwise pseudonymous identifiers (PPIDs). A counterparty may require PPIDs throughout its relationship with the same agent, for example to maintain state or reputation. Agents can present pairwise pseudonymous identifiers (agent PPIDs) in place of IDs, so long as PPIDs can still support identity escrow (ES-1) to the desired level of assurance. The agent must have at most one PPID for any given counterparty, and no two agents may share the same agent PPID. The system may also support a principal PPID, which is tied to the principal rather than the agent, such that it is the same for all of that principal's agents (PV-3) for any given counterparty. Actors must not be able to collude within the system to correlate an agent's or principal's activities unless expressly required and appropriately authorized.

PR-9: No directory contact. A counterparty can verify a presentation, including the validity of its claims (PR-3), using only the presentation itself and the information the issuing directory publishes publicly (DR-1), such as public keys or a list of each claim's validity status. It must not need to contact the directory to verify a given presentation.

PR-10: Presentation requests. A counterparty can request a presentation, or request claims that an initial presentation lacks, such as an authorization, the type of verified identity of the principal (domain, natural person, etc.), or the principal's identity assurance level. The counterparty may make subsequent requests, and the agent can then make subsequent presentations that include the requested claims according to its principal's policies. In requesting additional claims, the counterparty must be able to do so without directly contacting the directory or the principal.

4. Revocation

RV-1: Revocation. A registration or an individual claim can be revoked. A principal can revoke the registration of any of its agents at any time, and a delegating agent can do the same for its sub-agents (AG-3). The directory can revoke a registration per its published policies (DR-1). Revoking a registration revokes every claim issued for that agent, including its IDs. A principal or directory can also revoke a single claim it has attested (PR-4), such as an authorization; the agent's registration and its other claims are unaffected. A revocation does not undo the anonymity or unlinkability of presentations (PR-6, PR-7), though a directory may provide otherwise in its published practices (DR-1). The agent record is kept after revocation so identity escrow across previous presentations remains intact (ES-1).

RV-2: Revocation latency. For each kind of revocation in RV-1, the directory publishes a revocation latency (DR-1): the longest time after a revocation for which a claim it revokes can still verify. Once that time has passed, the claim does not verify (PR-3) for counterparties using the directory's current published information (PR-9). The latency is a bound that the directory's design enforces (DR-2), for example through how long claims stay valid or how often published revocation information is refreshed.

5. Identity Escrow

ES-1: Identity escrow. For any valid presentation sent by an agent, the verified identity of the agent's principal can be recovered only according to a process the directory publishes (DR-1), which includes how it responds to lawful requests. This holds for the whole retention period, including after revocation.

ES-2 (optional): Multi-party control. A directory may adopt technical or procedural measures that require several parties to cooperate to recover a verified identity, such that no single party, even the directory itself, is able to recover it alone.

6. Directory

DR-1: Published practices. A directory publishes the information counterparties need to verify its issued IDs and its practices, so counterparties can decide whether it meets their required trust framework. The practices can include supported proofing methods for principals and supported types of assurance levels (PV-1), which means of authentication it accepts from principals and agents (PV-5, AG-1), how it protects and retires the material it signs IDs with, how long IDs remain valid, the revocation latency of registrations and claims (RV-2), which information it stores and how, how long it retains records, organizational security practices, and its process for releasing information about principals and agents, including how it responds to lawful requests (ES-1).

DR-2: Adherence. A directory adheres to the practices it publishes under DR-1. For example, it keeps principal records, agent records, the means to connect a presentation to them, and records of any revocation for at least the retention period it publishes (ES-1).

DR-3 (optional, sector-specific): Conformity assessment. A directory may give counterparties and principals evidence about its adherence to published practices (DR-2). Possibilities include but are not limited to certification of conformity to a framework by a standards or certification body, interoperability testing results, an independent audit, certification against a sector's security standard, attested computations, and licensing or supervision by a regulator.

B Related Work

Many organizations are building protocols that standardize how AI agents share verifiable information with counterparties. This field is nascent and rapidly evolving; at a high level, existing work falls into four categories by function, which we survey here.

Identifying the agent. The Internet Engineering Task Force (IETF) is standardizing Web Bot Auth [87], which allows an agent to attach a short-lived digital signature created by the agent provider to any web request it sends. This allows counterparties to confirm the request came from the agent provider. Visa's Trusted Agent Protocol [88] and Mastercard Agent Pay [89] build on Web Bot Auth so that online stores can check that a card network has approved an agent's provider. AAuth [59] is more fine-grained than Web Bot Auth by giving each agent its own key. Standards from WIMSE [90] (Workload Identity in Multi-System Environments) enable a company to give individual running software programs a short-lived digital ID that identifies the program and company; the company's partners can then check that the ID is genuine [91]. Other approaches use public directories where anyone can look up an agent's details, such as its name and web address. For example, the Agent Name Service [58] lets anyone who controls a web domain register agents under that domain. Ethereum's ERC-8004 [31] standard lets anyone list an agent on the blockchain, where anyone can rate it.

Granting permissions. The Model Context Protocol [92] (MCP), a widely used standard for connecting AI agents to external services, has an optional step in which the service asks the principal to confirm that the agent should have specific permissions before granting them. Okta's Cross App Access [93] lets an employer grant these permissions centrally, so employees need not approve each connection themselves. Similarly, Sierra and Meta's Personal Agent Protocol [94] lets an agent use the principal's account at a business, for example to check on an order, with only the degree of access approved by the principal. The Agent Delegation Profile [95] lets an agent delegate tasks to subagents along with a signed record of what each subagent may do. For payments, Google's Agent Payments Protocol [96] (AP2) and Mastercard's Verifiable Intent [97] let a principal digitally sign the limits of a purchase in advance, such as "hiking shoes under $150." EMVCo [33] has drafted a framework, building on Verifiable Intent, for tracking such approvals over time.

Linking the agent to its principal. WorldID lets people prove they are a unique human without revealing their identity (proof of personhood), and its AgentKit [98] extension allows principals to register agents under their WorldID. Counterparties learn that an agent is backed by a unique human without learning that human's identity. Proof's x401 [99] lets a website request specific verified facts

about the principal, such as their age or legal name, which the principal can approve sharing. Under Skyfire's KYAPay [100], a company verifies the principal's identity and gives their agent short-lived tokens that refer to the principal by a different pseudonym at each website. These tokens can then be linked to the principal. GLEIF, which oversees Legal Entity Identifiers, has proposed that companies give their agents verifiable credentials that communicate the agent's company and permissions [101]. The Decentralized Identity Foundation's KYA-OS [102] lets a person or company digitally sign a credential attesting that an agent acts for them and specifying its scope, though the principal's identity does not need to be verified. By contrast, Privacy Pass [103] issues tokens that websites can verify but cannot trace back to whoever received them: a way to limit abuse without identifying anyone, and so without a route back to a principal. EMVCo's draft framework links an agent's payments to its principal: a card issuer signs a credential for the cardholder, the cardholder confirms what the agent may spend and its key, and the agent signs each purchase, so a payment can be traced from the agent to a cardholder verified by the issuer.

Managing agents within an organization. Products such as Microsoft Entra Agent ID, Amazon Bedrock AgentCore Identity, Google Cloud Agent Identity, Okta for AI Agents, Ping Identity for AI, and Alibaba Cloud Agent Identity work somewhat like human resources for AI: a company can create IDs for its agents, set permissions for what they can access, and monitor their activity [104–109]. These identities are recognized only within that company.

How AAF compares to existing work. Most of the related work discussed here aims to enable commerce and other agent activity rather than legal accountability. Furthermore, our framework prevents counterparties from learning who the principal is by default, while retaining the option for attribution post hoc. KYAPay is the closest existing work to AAF, though its draft specification does not currently satisfy every property in Appendix A; for instance, its tokens are bearer tokens rather than bound to the presenting agent's authentication (PR-2). However, many protocols satisfy many properties in Appendix A and should be considered at least partially AAF-compatible, with potential for full alignment as they continue to evolve. For instance, protocols that tie an agent to a payment card, such as EMVCo's draft framework [33], let card issuers serve much like directories for cardholders. In Appendix C, we reference existing protocols that could be included in (or used as precedent for) components of a fully AAF-compatible system.

Note that there are many other protocols besides those listed here [30].

C Implementation Options

This section surveys potential implementations of components of the framework, and briefly discusses their trade-offs with respect to the framework's design goals. This section is neither meant to be exhaustive, nor does it endorse specific implementations. Rather, its goal is to demonstrate the implementability of the framework using existing protocols and technologies. Requirement numbers refer to Appendix A.

Design Goals

- **Anonymity toward counterparties:** a counterparty learns nothing about an agent, its delegating agents, or its principal beyond the claims presented (PR-6).
- **Unlinkability:** counterparties cannot link an agent's presentations (PR-7).
- **Minimal directory knowledge:** the directory learns as little as possible about an agent's activity (e.g., PR-9).
- **Identity escrow:** a presentation can be used to recover its principal's verified identity under the directory's established process (ES-1).
- **Minimal revocation latency:** claims remain valid after revocation (RV-2) for as little time as possible.
- **Minimal overhead** for counterparties, agents and directories (e.g., PR-9).

1. Identity Proofing (PV-1, PV-2)

Option	Properties	Precedent or implementation
1. Verifying control of a domain	The verified identity is a domain, which can potentially lead to a natural person or legal entity through the domain's registrar	Domain-validated web certificates (CA/Browser Forum Baseline Requirements; ACME) [41, 110]
2. Human identity proofing	The verified identity is a natural person	NIST SP 800-63A (IAL1 to IAL3) [44]
3. Legal entity proofing	The verified identity is a legal entity, together with its authorized representative	Bank customer identification rules; Legal Entity Identifiers; extended-validation web certificates; commercial business-data providers [60, 111, 112]
4. Verifying control of an email or phone number	Moderate Sybil-resistance at the cost of a weaker verified identity	S/MIME Baseline Requirements, Section 3.2.2 (email); NIST SP 800-63B-4, out-of-band authenticators (phone) [45, 113]
5. Proof of personhood	The verified identity is a unique natural person, without disclosing who	Personhood credentials [70, 114]

Note: The method and assurance level of identity proofing can be expressed as a claim, for example with the `verified_claims` schema of OpenID Identity Assurance [115].

2. Authentication of Principals and Agents (PV-5, AG-1, AG-5)

Option	Properties	Precedent or implementation
1. Proof of possession of a key: the agent signs each request, including the counterparty's name and a time or a value the counterparty supplies	Needs no special hardware	DPoP; HTTP Message Signatures (as used by Web Bot Auth); key binding in SD-JWT; IETF WIMSE workload proof tokens [87, 116–119]
2. Proof of possession tied to the connection the presentation is sent over	The presentation cannot be relayed over another connection	Mutual TLS [120]
3. Key kept in certified hardware	The key is non-exportable, so it cannot be copied or stolen through the agent's software	AAL3 in SP 800-63B; FIPS 140-3 [40, 45]
4. Zero-knowledge proof of possession	The proof itself is unlinkable (section 3, option 3)	BBS signatures; anonymous credential systems [52, 121]

Note: Bearer tokens, such as OAuth access tokens, and shared secrets, such as API keys, can be reused by whoever obtains them, potentially undermining PR-2 [122, 123]. A bearer token that is restricted to one counterparty (its audience) and expires within minutes limits that to replay at that counterparty for a short window, as with OpenID Connect ID tokens (section 3, option 1) [124]. Whether that is enough depends on the use case, so counterparties can require proof of possession (options 1 to 4) where the risk warrants its added cost.

3. IDs and Unlinkable Presentations (PR-1, PR-2, PR-5, PR-6, PR-7)

Option	Properties	Precedent or implementation
1. On-demand issuance: the directory issues a presentation on demand, which is directly passed to the counterparty	Unlinkable if each presentation carries a fresh ID, though the directory learns the timing and claims of each presentation	OpenID Connect ID tokens with pairwise subject identifiers; SAML transient identifiers [124, 125]
2. Batch issuance: the directory issues many single-use IDs in advance, each bound to its own key, with claims disclosed selectively by salted hash	Unlinkable if each ID is used once; the directory learns batch timing only; the agent stores and replenishes batches	SD-JWT VC; ISO/IEC 18013-5 mdoc; batches of pseudonym certificates for vehicles (SCMS) [118, 126–128]
3. Anonymous credentials: the directory issues one credential and the agent derives a fresh proof from it for each presentation	Unlinkable by directory or counterparty, unless the directory colludes with counterparties to recover the principal identity for multiple presentations	BBS signatures; zero-knowledge proofs over standard credentials (Longfellow) [121, 129]

Notes:

- Options 2 and 3 follow the subscriber-controlled wallet model of SP 800-63C, with the directory as the credential service provider, the agent holding the wallet, and the counterparty as the relying party [130].
- Longfellow, as released, makes presentations untraceable even by the issuer. Using it under this framework would require designing identity escrow (ES-1) into the proof.

4. Authorizations (PR-4, RV-1)

Option	Properties	Precedent or implementation
1. Directory-attested standing authorizations, held at the directory and included with IDs	The directory learns the standing scope and nothing per transaction; no overhead at transaction time; revoked through the directory at the principal's request	OAuth scopes [131]
2. Directory-attested authorization for each transaction	The directory is contacted for each authorization and learns each approved transaction, usually with the counterparty; no authorization outlasts its transaction, so none needs revoking	OAuth Rich Authorization Requests [132]
3. Principal-attested authorizations, signed with a key the ID certifies	No interaction with the directory; needs a device or app for the principal that the agent cannot reach; revoked by expiry, or by revoking the agent's registration	Signed mandates (Agent Payments Protocol; NIST's proposed design tenets for agent identity) [96, 133]

Note: The options differ in how independent the issuer is from the agent. If the agent has access to the principal's authorization secrets, option 3 gives no more assurance than self-assertion.

5. Validity and Revocation (PR-3, PR-9, RV-1, RV-2)

Option	Properties	Precedent or implementation
1. Short-lived IDs with no status check	Revocation latency is the ID's lifetime; the directory learns issuance timing	Short-lived certificates; OAuth access tokens [131, 134, 135]
2. Stapled status: the agent fetches a fresh signed status and attaches it to the presentation	Revocation latency is the status' lifetime; the directory learns when the agent is active; repeating the status across presentations links them	OCSP stapling [136]
3. Published status list: the directory publishes a list of all valid IDs	Revocation latency is the list's refresh interval; the directory is not contacted about individual IDs; each single-use ID needs its own position in the list	IETF Token Status List; certificate revocation lists [137, 138]
4. Proof of non-revocation: the directory publishes the set of revoked credentials, and the agent proves its credential is not in it	Revocation latency is the set's refresh interval; the directory is not contacted about individual IDs; unlinkability is preserved	Cryptographic accumulators; zero-knowledge proofs [139]

Notes:

- A counterparty querying the directory about each ID, as in OCSP, does not satisfy PR-9, since the directory learns which counterparty checked which ID [140].
- Lower revocation latency generally costs more overhead, or tells the directory more about when an agent is active.

6. Identity Escrow (ES-1, ES-2)

Option	Properties	Precedent or implementation
1. The directory keeps a map from issued IDs to agent records	Low overhead, but the directory alone can recover verified identities	Issuance records kept by certificate authorities [41]
2. Each presentation carries a pointer that only the directory can decrypt	No per-ID map to keep, and works with anonymous credentials (section 3, option 3); as in option 1, the directory alone can recover verified identities	Group signatures with an opening authority; inspection in anonymous credential systems (Idemix) [141, 142]
3. The map or the decryption key is split among several parties, a threshold of whom must cooperate (ES-2)	No single party can recover the verified identity; higher overhead for recovery	Threshold cryptography; SCMS [128, 143]

7. Agent PPIDs (PR-8)

Option	Properties	Precedent or implementation
1. The directory derives the identifier from the agent record and the name of the counterparty, or group of counterparties, at issuance	Low overhead, but the directory learns the counterparty	Pairwise and shared pairwise pseudonymous identifiers in SP 800-63C; OpenID Connect pairwise subject identifiers [124, 130]
2. The agent provably derives the identifier from a directory-attested secret bound to its agent record and the name of the counterparty or group	Higher overhead, but the directory is not contacted	Scope-exclusive pseudonyms in anonymous credential systems; BBS pseudonyms; Restricted Identification in the German eID [144–146]

Note: Similar options can support a principal PPID.

8. Directory Key Management (DR-1)

Option	Properties	Precedent or implementation
1. Signing keys kept in certified hardware	Non-exportable keys cannot be copied out of the directory's systems	FIPS 140-3; CA/Browser Forum Baseline Requirements [40, 41]
2. A root key that certifies online issuing keys	A compromised issuing key can be revoked and replaced without changing the root key that counterparties trust	Certificate authority hierarchies (CA/Browser Forum Baseline Requirements) [41]
3. Scheduled retirement of signing material	Bounds how long IDs signed with retired or stolen material still verify; counterparties must more frequently retrieve the directory's published information	Certificate lifetimes and key rollover

9. Arrangements Among Directories (PV-4, DR-3)

Option	Properties	Precedent or implementation
1. Trust list: one party publishes which directories belong to a group and under which policies	Counterparties learn a directory's group membership and rules from one source	OpenID Federation 1.0; EU trusted lists of trust service providers; SP 800-63C federation [130, 147, 148]
2. Duplicate check across directories before a principal record is created	One verified identity has one principal record across the group; directories learn about principals in other directories	Legal Entity Identifiers, whose issuers check each new identifier against all others before issuing [149]
3. Audit or certification against a common standard	Provides evidence that directories are adhering to standards and their published practices	WebTrust and ETSI audits of certificate authorities [150, 151]

D Security Analysis

If agent IDs are used to gate access to valuable services, such as bank accounts, sensitive materials, or governmental processes, the identification system will likely come under attack by malicious actors. Widely deployed identity systems, including TLS, DNS, and WebPKI, have made significant progress in solving key distribution and revocation, through decades of operation at scale. AAF implementations can start from those systems' design choices. Below we outline some threat models that any implementation of agent IDs must address.

Spoofing. One threat model is *spoofing*: falsifying identifiers or records to gain illegitimate access or evade accountability [44]. A principal may misrepresent its true identity to a directory, either to gain access to a service they are not entitled to use or to prevent harmful actions from being traced back to them; a malicious actor might take over a legitimate principal's directory account or use stolen identity documents to open one, and then register agents under the compromised principal's identity; an agent may steal another agent's authentication secrets to misrepresent itself to counterparties.

Directories can mitigate identity fraud by using robust mechanisms to verify principals' identities. At registration, directories can use assurance methods along varying degrees of rigor depending on the context, from email and phone verification, to live identity verification, to personhood credentials [70, 86, 100]. Directories can also re-verify principals for high-risk actions, notify principals of new agent registrations, and perform anomaly detection. Each credential should state how the principal was verified, so that counterparties can apply appropriate thresholds of rigor.

Directory vulnerabilities. A compromised or careless directory might issue IDs for agents whose principals do not exist, or intentionally alter records for personal benefit [152]. These risks can be mitigated through governance of directories, such as independent audits, baseline operating requirements, and publicly auditable issuance logs. See Section 3.1 for more on directory governance.

Agent hijacking. Both identity spoofing and governance failures are existing threat models from the pre-agentic internet. As autonomous systems, agents also introduce a novel threat vector: agent traps [66]. With techniques like prompt injection [64, 65], an adversary might manipulate an agent's reasoning and influence its behavior. Because a hijacked agent signs presentations using its own legitimate key, holder binding (PR-2) alone cannot distinguish an authorized action from a prompt-injected one. Trust in agents as they transact thus also requires robust security measures for the AI agents themselves.

None of these threats can be eliminated completely. However, like most risks within modern identity infrastructure, they can arguably be reduced to a practical level.