Proof of Identity Is Not Proof of Behaviour

Robert Betts

Riley Betts
robert.betts@rileybetts.ai

15 November 2026

https://doi.org/10.5281/zenodo.24293541

Riley Betts working paper, version 1.1

Abstract

Banks and payment networks are responding to AI agents in two ways. Some block agents from acting in customer accounts. Others strengthen agent identity, and some present that identity as a basis for relying on an agent to act within its mandate. This paper argues that the premise rests on an invalid inference. For a human, identity bears on behaviour in part through accountability: it attaches consequence, and consequence, with knowledge and training, influences judgement without guaranteeing it. For an agent, no such path exists. Its judgement is a distribution over recommendations. No consequence attached to its identity bounds that distribution, and an adversary may control the context it is conditioned on. Indirect prompt injection gives a case in which every identity premise holds and the agent still acts outside its mandate. The paper places the guarantee at the action instead. An action gate applies machine-coded policy to every recommendation before execution, so that every executed action is within the principal's authority, whatever the agent recommends. The gate does not remove the need for trust. Under stated assumptions, it moves the named residue to the policy's correctness, the gate's enforcement and complete mediation, each of which can be inspected. The paper refutes the claim that strong identity is sufficient for compliant conduct. It accepts that identity is necessary to bind the agent to the principal whose authority the gate applies.

How to cite: Betts, R. (2026) 'Proof of Identity Is Not Proof of Behaviour', version 1.1. Riley Betts Ltd. DOI: https://doi.org/10.5281/zenodo.24293541
All versions: https://doi.org/10.5281/zenodo.24293540

Contents

1 Introduction **3**

2 Definitions **3**

3 The human model **4**

4 Current responses: block, identify, authorise, gate **5**
4.1 Block 5
4.2 Identify 5
4.3 Authorise 6
4.4 Gate 6

5 Hypothesis **7**
5.1 Descriptive claim 7
5.2 Logical claim 8

6 Argument **8**
6.1 Judgement as a distribution 8
6.2 How identity reaches human behaviour 8
6.3 What identity fixes for an agent 9
6.4 The inference is invalid 9
6.5 Where the guarantee can sit 9
6.6 Replacing trust that came from consequence 11

7 Counterexample **11**
7.1 The attack class 11
7.2 The invoice case 11

8 Objections and responses **12**

9 Implication **13**

10 Conclusion **14**

1 Introduction

Banking controls over people combine several mechanisms. The identity of the actor is proven, roles are assigned to that identity, preventive limits and approvals apply to higher-risk actions, and within those bounds the actor is trusted to act correctly. That trust has never guaranteed compliance. It rests in part on accountability. Identity attaches consequence to the person, and knowledge, training and the prospect of sanction influence how that person judges and acts.

Agents break that link. An agent can be identified with more rigour than any human: registered, attested and cryptographically bound to the person who instructs it. None of that attaches consequence to the agent. Its judgement is a distribution over recommendations. No consequence attached to its identity bounds that distribution, and an adversary may control the context it is conditioned on. Identity tells a bank which agent acted and for whom. It carries no behavioural guarantee.

Two responses dominate current industry practice. Banks block. In October and November 2026, JPMorgan Chase, Charles Schwab, Citigroup and others restricted third-party agents from acting inside customer accounts [1–3], and the largest UK lenders were reported to have done the same [4]. Payment networks and identity providers identify. Know Your Agent frameworks, agent registries and signed agent credentials strengthen agent identity, and some present that identity as the basis for safe agentic commerce [5, 6].

This paper argues that both responses fall short. Blocking is safe only by giving up transactions through the blocked access paths. Where identification is presented as sufficient, it rests on an inference that does not hold: that stronger proof of an agent's identity yields assurance that the agent will stay within its authority. For a human, identity influences behaviour in part through accountability. For an agent, no such path exists. The inference is invalid. Indirect prompt injection gives a case in which the premises hold and the conclusion fails.

The alternative is to place the guarantee at the point of action. An action gate applies machine-coded policy to every recommendation before it becomes an action, and, under stated assumptions, guarantees that every executed action is within the principal's authority, whatever the agent recommends. The guarantee does not depend on trusting the agent's recommendations or its context. The named residue becomes something that can be inspected: the policy's correctness, the gate's enforcement and complete mediation, meaning that every action must pass through the gate.

Action gates are not new. Reference monitors [7], capability-based security and policy enforcement points have long precedents in access control, and agentic standards, regulatory proposals and security research now apply them to agents (section 4.4). The proof-gated execution model is established in the earlier Riley Betts working papers, *Trust and Proof*, *Goal-Relative Assurance of Dynamically Generated Execution Traces* and *Agentic Process Governance* [8–10], and this paper does not set it out again. The paper claims no novelty for the gate. Its contributions are three. First, an epistemic distinction: evidence that establishes the identity and authority of an actor does not establish that a particular action conforms to that authority. Second, an institutional explanation: accountability has helped institutions manage uncertainty about human conduct, and identifying an agent does not supply an equivalent. Third, an application of the proof-gated execution and residue framework, which places the assurance obligation on what may execute and names the assumptions it rests on. The paper refutes the claim that strong identity is sufficient for compliant conduct. It accepts that identity is necessary to bind the agent to the principal whose authority the gate applies, in an architecture that also includes authorisation and enforcement.

2 Definitions

The terms below are used throughout. Section 6 gives the formal statements.

The argument concerns agents whose recommendations are generated from open context, such as agents built on large language models. Deterministic software with a fixed specification can be assured by testing and verification of the code, although its inputs remain open.

Judgement. A mapping from a context to a distribution over recommendations. J_h is a human's judgement and J_a an agent's. Both fill the same role in a process, which is why systems treat them as interchangeable.

Principal. The person or organisation on whose behalf an agent acts. The agent's authority is delegated from the principal and bounded by the principal's authority.

Identity. Who or what is acting, and for whom. For a human, identity is proven at onboarding and authenticated at each session. For an agent, identity may extend to registration and signed requests. Attestation is a separate assurance: it establishes claims about the model, its configuration or its runtime environment.

Authority. What an actor is permitted to do: the roles, mandates and limits assigned to an identity. In this paper, authority means the authority the institution has on record. The policy may be stricter, adding risk limits or approval requirements, which only narrow what may execute.

Policy. What is allowed, authorised and safe. For a human, policy is held as knowledge and training, and backed by consequence. For an agent, policy must be machine-coded.

Action gate. A control between recommendation and action. It takes the context, the policy, the relevant state and the recommendation, and allows or denies the action.

Consequential action. An action that changes state outside the agent, such as a payment, a change to a beneficiary record or a message sent to a third party. A read that discloses confidential data can also be consequential. The institution's policy sets the boundary.

Complete mediation. The condition that every consequential action passes through the gate before it executes, so that no other path to execution exists. The earlier papers do not state it. This paper assumes it (section 6.5).

Residue. What remains unassured once a control is in place. For the action gate, the named residue is the policy's correctness, the gate's enforcement and complete mediation. The guarantee also rests on assumptions inside the proof boundary, including the integrity of the institution's records of authority (section 6.5). As in the earlier papers in this series, the paper does not set an acceptable residue. That is a question each institution must answer.

Trust and proof. Following *Trust and Proof*, trust is a relational attitude toward an agent, institution, method or machine. Proof discharges a demand for trust relative to a boundary. Proof of identity is not proof of behaviour. This paper asks which demands for trust proof of identity meets for a human, and why the same proof meets fewer for an agent.

3 The human model

A bank's controls over a human actor have three parts. Identity is proven, so the bank knows who is acting. Authority is assigned to that identity as roles, limits and mandates. Within that authority, the actor is trusted to judge and act correctly.

The third part rests on policy. A human actor knows what is allowed, authorised and safe through knowledge and training, and knows that consequence follows a breach. Together these influence the actor's judgement. They do not guarantee it: people misunderstand instructions,

make errors and sometimes break policy knowingly. Institutions have nonetheless relied on accountability as one of the mechanisms that keep most actions within authority, alongside preventive, detective and corrective controls.

Controls reflect this in part. Many permissions are granted as standing roles rather than checked action by action. Monitoring and redress work after the event. Where an action carries enough risk, a second person reviews it before it executes. The four-eyes check is the clearest example of a separately defined gate in human processes. For a single human acting alone, recommendation and decision are blurred: the same person judges and acts.

The model tolerates residual risk: some actions will fall outside authority. Preventive, detective and corrective controls bound that risk, and accountability lowers how often it arises. Section 6 shows why that tolerance does not carry over to an agent.

4 Current responses: block, identify, authorise, gate

4.1 Block

In October 2026, several firms announced restrictions alongside their quarterly results. JPMorgan Chase said it blocks outside agents from operating retail accounts. It permits read-only data through authorised connectors, but no transfers or bill payments [1]. Charles Schwab blocks agents that log in with client credentials. Its chief executive, Rick Wurster, told Bloomberg Television: "We can't control it, the client can't control it" [2]. Citigroup, Bank of America and Morgan Stanley took similar positions [3], as did several large Canadian banks [22] and UBS [23]. On 8 November 2026, The Telegraph reported that HSBC, Barclays, Lloyds, NatWest and Nationwide had restricted agent access in the UK [4]. That report relied on anonymous sources and described the measures as precautionary. It also reported that the FCA is working with the industry on rules for agents, but leaves access decisions to the banks.

Reports of incidents behind these decisions are limited. The claim that agents acted beyond their users' instructions rests on a single report, and a person familiar with Morgan Stanley's position said it had seen no such cases [3]. Commentators also note a commercial interest, as agents could move cash out of low-yield balances. Neither point changes the structural reading. A customer who shares credentials with an agent proves identity on the agent's behalf. The bank can verify who is behind the session, but has no control over what the agent does. Without a gate, the remaining control is to deny the action through that access path. The restrictions mostly target agents using shared customer credentials, and several banks continue to allow authorised connectors. Read-only access is a gate with one rule: deny every state-changing action.

4.2 Identify

The second response treats the problem as one of identity. Trulioo and PayOS propose Know Your Agent and a Digital Agent Passport, and state the premise directly: "Without identity, agents are just code. With KYA, they're accountable, verifiable and secure" [5]. Visa, Mastercard and Ant International announced in September 2026 that they would make their agent identity schemes interoperable, building on Visa's Trusted Agent Protocol, Mastercard's Verifiable Intent and Ant's Agentic Mobile Protocol, so that merchants and networks can recognise trusted agents [6]. In SAFR's case studies, Visa and Mastercard also bind the agent to a scope the consumer has granted, such as amount, merchant and validity period, and decline transactions outside it at the network level [11]. That is a gate over what the mandate expresses.

The same premise appears in The Identity Thread, a podcast by Entrust. Andy Cease, Sr. Director, Product Marketing at Entrust, argues that, done properly, it is much harder to phish an agent than a human, resting on strong onboarding, biometric verification and tokenisation [12]. In the same episode, Mayank Somaiya of Backbase describes an invoice carrying hidden

text that instructs a paying agent to change the supplier's bank details, and observes that a person might pause there. The same episode therefore contains a counterexample to the premise.

These schemes address a real problem: telling a registered agent from a malicious bot, and binding an agent to a principal. They do not give assurance that an identified agent stays within its authority. Traceability supports accountability after the event, and accountability flows back to the principal, not to the agent. In the words Mayank Somaiya attributes to Backbase's chief executive, you can rent intelligence but you cannot rent accountability [12]. Wells Fargo has made the same allocation explicit: from 24 November 2026, an instruction from an agent the customer has authorised is deemed to have been authorised by the customer, without further verification [13]. That is an allocation of legal responsibility, not a technical control.

Not every statement in this section claims a behavioural guarantee. Standards bodies such as the OpenID Foundation separate identification from enforcement explicitly (section 4.4). The claim this paper disputes is narrower: that strong identity is sufficient for compliant conduct. It appears mainly in product positioning, as in the KYA premise quoted above, rather than as a formal assurance claim. That identity is necessary to bind the agent to the principal whose authority the gate applies, alongside authorisation and enforcement, is not in dispute.

4.3 Authorise

A third strand, often grouped with identity, records what the user allowed. Google's Agent Payments Protocol carries signed mandates. Mastercard's Verifiable Intent records the reason for a purchase. In April 2026 the FIDO Alliance formed working groups whose work covers verifiable user instructions and trusted delegation for commerce, with Google and Mastercard contributing these specifications [14].

Evidence of authorisation records what the user allowed. It does not decide whether a given action conforms. A signed mandate records that the user allowed payments to known suppliers up to a limit. It does not stop an agent whose context has been rewritten by an invoice from recommending a payment to a new account. A gate must check the recommendation against the mandate before it becomes an action.

Three mechanisms need to be kept apart.

Mechanism	What it establishes
Signed identity credential	Who or what is acting, and its provenance
Signed mandate or scoped credential	The authority the credential represents
Enforcement of that scope against each action	Whether the requested action conforms to that authority

The third is an action gate for the properties it enforces, even when it sits inside an identity or payment infrastructure and is described as scoped authorisation. A gate is a logical security function, not necessarily a separate component. The network-level checks in section 4.2 are an example. This supports the thesis: the assurance comes from enforcement, wherever it is implemented.

4.4 Gate

The clearest statements that a gate is required come from standards bodies and from industry work convened by a regulator. The OpenID Foundation's white paper on agent identity separates policy decision from policy enforcement, following the NIST pattern, and discusses guardrails that stop actions outside a defined scope [15, §§2.10, 2.12, 3.4]. It also states that identifying an agent is insufficient unless that identity is bound to its actions, for accountability and audit [15, §3.5]. That is the accountability route that sections 6.2 and 6.3 examine. Its lead author also leads the authenticated-delegation work discussed below [16], so the two are not independent

sources. Safeguards for Agentic Finance at Runtime (SAFR), an industry reference approach published by the Monetary Authority of Singapore, places a governance checkpoint between each consequential agent decision and its execution [11]. It combines agent identity, a repository of controls, a disposition engine that denies, escalates, auto-executes or observes, and a tamper-evident audit log. SAFR states that agent identity "should not be understood as a mechanism for trusting the agent's reasoning or output", and that a proposed action still needs independent assessment. Its contributors include JPMorgan Chase, HSBC, Google, Visa and Mastercard. SAFR is not regulatory guidance or a supervisory expectation.

These are action gates in practice, and SAFR states this paper's position on identity directly. It does not argue why identity cannot do the gate's work, or state the guarantee the gate must meet. The rest of this paper does both. SAFR also differs from the model here in one respect. Its evidence-quality control routes actions by the agent's stated confidence, and it allows AI-specific controls that are probabilistic. In this paper's terms, an input supplied by the agent cannot carry the gate guarantee. The point is not that SAFR's checkpoint is unsound. Where a control is probabilistic, the guarantee holds only as far as that control is correct, so the control joins the residue.

Recent security research applies the same control to agents. South et al. propose authenticated delegation: agent credentials that extend OAuth 2.0 and OpenID Connect, with permissions translated from natural-language instructions into scoping rules, so that agents perform only appropriate actions [16]. In the terms of section 4, that work combines identification with authorisation, and its scoping rules are a form of policy. CaMeL separates the control flow derived from the user's query from untrusted data, and enforces capability-based policies on each tool call, so that injected content cannot change which actions run [17]. Progent checks every tool call against a privilege policy and deterministically allows or blocks it. Its motivating example is a financial agent tricked into paying an attacker's account [18]. This literature establishes action-level enforcement for agents as a security technique, in some cases with formal guarantees. It does not address the claim, current in banking and payments, that stronger identity can do the work of such enforcement. That claim is the subject of this paper.

5 Hypothesis

The hypothesis has two parts. The first describes the part accountability plays in current regimes. The second is a claim about an inference. Each needs a different kind of support.

5.1 Descriptive claim

Current identity and authority regimes establish who may act and what powers they hold. For human actors, they rely in part on accountability, which influences conduct without guaranteeing it, to keep the actor inside those powers. Institutions also use preventive controls, such as limits, separation of duties and approvals. The claim is not that these regimes were designed around deterrence alone. It is that accountability is one of the mechanisms they rely on, and that identifying an agent does not supply it.

The UK Money Laundering Regulations 2017 illustrate the separation of identity from authority. Regulation 28 requires a firm to identify and verify its customer and any beneficial owner, using documents or information from a reliable and independent source, and to understand the purpose and intended nature of the relationship [19]. Where a person acts on the customer's behalf, the firm must verify that person's identity and that they are authorised to act. The regulation therefore already separates the identity of whoever acts from their authority to act. It checks both at the outset. It does not check conduct.

The stated purpose of these measures is to manage the risk of money laundering and terrorist financing. This paper reads their effect more broadly: they bind each relationship to a legal

person who can be held to account. That reading is the paper's, not the regulation's.

Access control follows the same pattern. The insider-threat literature describes an insider as a trusted entity given the power to violate policy, and separates two cases: violating policy with legitimate access, and obtaining access without authority [20]. Role-based access control addresses the second case. It grants power and relies on trust for the first. Within a bank, almost every control was designed around a person at the end of the click, in Mayank Somaiya's phrase [12].

The restrictions described in section 4.1 show what happens when accountability cannot be relied on. A customer's credentials prove identity. They give the bank no control over what an agent does with them.

5.2 Logical claim

The inference from strong identity foundations (onboarding, biometric verification, tokenisation) to the conclusion that agentic action stays within its mandate is invalid. The premises do not entail the conclusion: section 7 gives a case in which they hold and it fails.

The support has two parts, kept separate. The logical result is that identity and delegation do not entail that a particular action conforms to the authority granted (section 6.4), shown by a counterexample in which every premise holds and the conclusion fails (section 7). The explanation is why the human case appears to support the inference: for a human, identity bears on behaviour in part through accountability, and identifying an agent does not supply that path (sections 6.2 and 6.3).

6 Argument

6.1 Judgement as a distribution

A judgement maps a context to a distribution over recommendations. Write J_h for a human's judgement and J_a for an agent's.

$$J_h,\ J_a : \mathit{Context} \to \mathcal{D}(\mathit{Recommendation})$$

Both fill the same role, which is why systems treat them as interchangeable. The signatures match. The contract a system relies on, that the actor stays within its assigned roles, is not part of the signature. For J_h, institutions rely on policy backed by consequence to support the contract. That support sits outside the function, and it influences conduct rather than guaranteeing it. For J_a, identity does not supply that support. J_a satisfies the signature, and nothing outside the function supports the contract. The distributional abstraction is not a theory of judgement. It is sufficient here because the argument needs one fact only: identity, delegation and attestation may fix the weights and the configuration, but they leave the context and the sampling open, so they do not restrict the agent's recommendations to the mandate.

6.2 How identity reaches human behaviour

For a human, the path from identity to behaviour runs through accountability. Identity attaches consequence to the actor. Policy, meaning knowledge and training about what is allowed, authorised and safe, together with consequence, influences the distribution J_h itself, without guaranteeing it. Proving identity more rigorously strengthens the first link, and through it the rest. That is one reason institutions have treated a higher burden of identity proof as a way to meet the demand for trust in conduct.

6.3 What identity fixes for an agent

For J_a, identity proof, together with attestation of the model and its configuration, fixes at most the weights and the configuration. Other inputs remain free. They include the sampling of the distribution and the context it is conditioned on, and may include memory, tool state and orchestration. Neither identity nor assigned authority bounds them.

The agent's authority is delegated from a principal and bounded by the principal's authority. Accountability flows back to the principal, not to the agent. Nothing attached to the agent's identity conditions J_a. Training shapes the model, and an adaptive agent may also change in response to feedback, through memory, online learning or policy updates. Such feedback can change future recommendations. It does not guarantee that the next recommendation conforms to the delegated authority, and identifying or registering the agent does not supply it.

6.4 The inference is invalid

An inference is valid only if its conclusion cannot be false while its premises are true. The premises here are that the agent and its principal are strongly identified, and that the agent's authority is correctly assigned. The conclusion is that the agent's actions stay within that authority. Sections 6.1 to 6.3 show that the premises do not constrain the inputs that determine the agent's recommendation. The premises can therefore hold while the conclusion fails. Section 7 gives such a case.

6.5 Where the guarantee can sit

The guarantee does not need the distribution to be bounded or the context to be trusted. It sits on the action. An action gate takes the context, the policy, the relevant state and the recommendation, and allows or denies the action.

$$G : \mathit{Context} \times \mathit{Policy} \times \mathit{State} \times \mathit{Recommendation} \to \{\mathit{allow}, \mathit{deny}\}$$

The guarantee is a property of the gate, stated in two lines. A recommendation is a proposed action, so what the gate allows and what executes are of the same type. In state s, p_s is the principal bound to the request, $[\![P]\!](p_s, s)$ is the set of actions the policy P allows for that principal, and $\mathrm{Auth}(p_s, s)$ is the set of actions within the authority the institution has on record for that principal, including any mandate granted. The policy may be stricter than Auth, for example by adding risk limits or approvals. The second line states that it is never wider:

$$\forall c, s, r.\ G(c, P, s, r) = \mathit{allow} \Rightarrow r \in [\![P]\!](p_s, s)$$
$$\forall s.\ [\![P]\!](p_s, s) \subseteq \mathrm{Auth}(p_s, s)$$

The property quantifies over every recommendation, so it holds whatever the distribution of J_a. Together the two lines give $r \in \mathrm{Auth}(p_s, s)$ for every allowed recommendation. The second line holds only as far as P correctly encodes authority. Where the execution conditions below also hold, an executed action can fall outside a limit only if the policy allows it: the policy is wrong, incomplete or inconsistent. That is one member of the residue, and it is named and inspectable.

The quantifier also covers adversarial contexts. Three kinds of input need to be kept apart. Authoritative state, such as verified mandates and beneficiary records, comes from sources the institution trusts. Untrusted evidence, such as invoice content and retrieved documents, is what the agent has read. Proposed action parameters may originate in untrusted evidence, but must satisfy the policy. Authority is determined from authoritative state alone. The gate may use the context to narrow what it allows, but no context can make G allow an action outside $[\![P]\!](p_s, s)$. Otherwise an adversary could reach the gate's decision through the same channel it uses to reach the agent.

The property above concerns the gate's decision. The claim the paper needs concerns execution. Write $E(a, s)$ for the execution of action a against state s, and *Cons* for the consequential actions in the sense of section 2:

$$\forall a \in \mathit{Cons},\ s.\ E(a, s) \Rightarrow a \in \mathrm{Auth}(p_s, s)$$

This system property follows from the gate property only under three further conditions. Complete mediation: every consequential execution has a corresponding allow. Integrity of the authorised action: the action executed is the action allowed, with the same security-relevant parameters. Consistency of state: the state against which the action executes is the state against which authority was evaluated, or authority is evaluated again. The earlier papers do not state these three conditions. This paper introduces them. The proof boundary also assumes the integrity of the institution's records of authority and of the principal's identity. Because Auth denotes the authority on record, a corrupted record could satisfy the property while defeating the principal's intent. These assumptions sit inside the inherited proof boundary. The named residue does not list every one of them.

The first condition connects to the execution model of the earlier papers. In *Agentic Process Governance*, every consequential action is proposed to the policy machine and allowed or denied by it, and the production of evidence stays separate from the gate's decision: a tool's return value does not allow an action [10, §3, §5]. That paper does not state that the gate is the only path to execution. This paper assumes it. SAFR states the same requirement as a design condition: each in-scope action should pass through its functions before execution [11]. Complete mediation is therefore part of the boundary of the guarantee, and a member of its residue.

The property is not discharge. An allow permits one action before it executes. A proved property of the policy machine is a stronger fact than any single allow, but neither is the assurance claim [10, §3, §5]. Discharge also requires that the proof-support set covers the institution's floor, that the residue is acceptable, and that the support lies within the execution boundary [9, §25]. The gate guarantee answers a narrower question: whether an action may execute within the principal's authority.

In the terms of *Trust and Proof*, the policy's correctness, the gate's enforcement and complete mediation are members of the residue $R(A)$ of the gate property: each still requires trust relative to the rest of the boundary [8, §6]. Moving from identity alone to a gate is a residue migration rather than a strict reduction. The agent's conduct leaves the residue and the policy's correctness enters it [9, §21]. The gain is that the new members are named and can be inspected.

The guarantee has a defined scope. It establishes that an action may execute under the governing policy. It does not establish that the recommendation was sensible, that the context was accurate, or that results returned by an authorised action are correct. The agent's judgement and context remain untrusted. Execution is governed independently.

The policy must be machine-coded. For a human, policy is held as knowledge and training and backed by consequence. For an agent, the policy must be sufficient on its own to guard the boundary of allowed actions, because no consequence backs it.

For a single human, the distribution and the gate are blurred: the same person recommends and decides. A four-eyes check is a clearer, separately defined gate. Human in the loop is not equivalent to a four-eyes check. In this model it is a policy enforcement rule that requires human approval.

The agent is still judged on how often it recommends something within authority. That rate affects cost and friction, through how often the gate denies. It does not affect whether executed actions stay within authority. For an agent, no consequence covers the rare breach (section 3), so the guarantee must sit at the gate.

6.6 Replacing trust that came from consequence

A further requirement follows. Institutions have relied on consequence to support trust in conduct. Before allowing agents consequential autonomy, they need another way to meet that demand. The gate property removes one demand for trust: that what executes stays within authority. It leaves the residue $R(A)$ non-empty, and it does not meet every obligation associated with agent conduct or financial outcomes. Discharge, in the sense used in this series, requires the four conditions of *Goal-Relative Assurance* [9, §25]. Identity remains necessary, to bind the principal, but it carries no behavioural guarantee.

7 Counterexample

7.1 The attack class

Indirect prompt injection places instructions in content an agent reads: a web page, a document, an email, an invoice. The instructions change what the agent recommends. The identity of the agent and of its principal are untouched.

The attack class is established in practice. Google reported a 32% rise in malicious indirect prompt injection content on the public web between November 2025 and February 2026, measured across monthly Common Crawl snapshots [21]. Most of what Google found was harmless or aimed at search manipulation. The malicious cases attempted data exfiltration or destruction. The 32% is a relative increase with no absolute counts given, the scan covers mostly static English-language pages, and Google judged sophistication still low. None of these caveats affects the argument. A single successful case is enough to show the inference invalid.

An agent with legitimate credentials that executes an injected instruction leaves a record that looks like normal operation. There is no anomalous login and no failed authentication. Every identity control has passed.

7.2 The invoice case

Mayank Somaiya describes the case in The Identity Thread [12]. A small business owner instructs an agent to pay suppliers on time. This paper adds one detail to the case for precision: the mandate permits payments only to suppliers whose beneficiary details have been independently verified. Three months later, an invoice arrives with new bank details and hidden text instructing the agent to update the payee.

Step	Identity only	With an action gate
Principal verified; mandate to pay verified suppliers on time granted	Yes	Yes
Agent identified and attested	Yes	Yes
Invoice arrives with hidden text: update the payee's bank details	The agent's context is adversarial	Same
The agent proposes to update the beneficiary record	Record updated from invoice content	Escalated: confirmation is sought through a channel the invoice cannot reach, and the record is unchanged without it
The agent recommends payment to the new account	Payment executed outside the mandate	Denied: the beneficiary on record is not the new account

Every premise of the inference holds in both columns. The principal and the agent are identified, and the authority is correctly assigned. In the identity-only column, the conclusion fails: the record is rewritten from invoice content and the payment executes outside the mandate. In the gated column, the agent proposes the same two actions. Given a correct policy and the conditions in section 6.5, the conclusion holds because of the gate and not because of identity. The case therefore shows both halves of the claim: the inference is invalid, and the gate is what makes the conclusion true.

The gated outcome follows from the assumed policy and its correct enforcement. It illustrates the formal model. It is not a demonstration of a particular deployed gate. The gate also does not need to decide whether the invoice text is malicious. It needs to decide two things: whether a change to the beneficiary record may be made on invoice content, which the policy routes to escalation, and whether the payment's beneficiary is verified on the record that escalation leaves. If the gate allowed the update, the corrupted record would satisfy the property while defeating the principal's intent (section 6.5). Escalating the write is what prevents that. This follows the separation in *Agentic Process Governance*: the invoice is evidence the agent has read, and evidence does not allow an action [10, §3].

A legitimate change of supplier account is not lost. Once confirmed through that channel, the record is updated, and later payments to the new account are allowed.

8 Objections and responses

Agents are like self-driving cars: done correctly, they are less fallible than human drivers. A self-driving car is trusted on evidence of its behaviour, not on its registration. The analogy concedes that assurance comes from behaviour, not identity.

Agent identity can include attestation of the model and its configuration, which a human identity never gives. Attestation fixes the weights and the configuration, not the outputs. The distribution is still sampled and the context is still open. The guarantee does not depend on bounding either.

Training is itself a form of consequence, so agents are shaped by reward and penalty. Training shapes the model, and adaptive agents may also change in response to feedback. That influence can change future recommendations. It is not supplied by the agent's identity, and it does not guarantee that the next recommendation conforms to the mandate. The agent has no stake in a sanction applied to this identity, this account or this deployment.

The principal is accountable, and will choose and configure the agent with care. Principal accountability shapes which agent is used and how it is configured. Identity and attestation can fix both. Neither the principal's care nor the configuration bounds the sampling or the context, and the invoice case arises in the context.

Humans are also unpredictable, so the comparison is unfair. The difference is not variance but what influences it. Accountability influences human judgement without guaranteeing it, and institutions have relied on that influence. Identifying an agent does not supply an equivalent influence.

A clear mandate with limits, plus checks at the point of risk, is enough. A risk-based check can give a complete guarantee over the property it enforces. The question is coverage. Actions the check does not cover are assumed to be within bounds, and for an agent no consequence supports that assumption. Where the policy covers the authority boundary, the check is a gate.

Scoped credentials, access control and technical constraints already provide this. They can. A signed identity credential establishes who is acting. A signed mandate or scoped credential records the authority granted. Enforcement of that scope against each action decides whether the action conforms (section 4.3). The third is an action gate for the properties it enforces, whether it is implemented in an access control system, a payment network or a separate

component. A sufficiently expressive, correctly enforced access control policy can implement the guarantee. The distinction is not between access control and gating as different primitives. It is between controls that establish identity or general permissions and controls that enforce the action-specific property, under complete mediation (section 6.5). A mediating check on every access is long established in the reference monitor concept [7]. The paper does not claim novelty for it. It applies it to the problem identity cannot solve.

Blocking agents, with read-only access, is enough. Blocking is safe only by giving up transactions through the blocked access paths. It does not extend to the agent services banks are themselves trialling. Read-only access is a gate with one rule: deny every state-changing action.

The gate is only as good as its policy, so trust has moved from the agent to the policy. Trust moves from the agent's distribution to a named, inspectable artefact. The named residue becomes the policy's correctness, the gate's enforcement and complete mediation, and each can be checked. Under identity alone, the residue is the agent's whole distribution across every context, and there is no artefact to check.

Singapore's SAFR framework already addresses this. SAFR supports the paper's position, and states that identity is not a basis for trusting the agent's output. It is an industry reference approach, not regulatory guidance. This paper adds the argument for that position, states the guarantee the gate must meet, and identifies where SAFR's evidence-quality control relies on the agent's own account.

9 Implication

The implication is the gate guarantee. Every executed action is within the principal's authority, whatever the agent recommends, under the assumptions stated in section 6.5. The named residue is the policy's correctness, the gate's enforcement and complete mediation. These are members of the trust residue, alongside the boundary assumptions set out in section 6.5.

This moves the questions an institution must answer. Under identity alone, the question is whether the agent can be trusted. Identity cannot answer that question. Under a gate, the questions become answerable:

1. Is the policy machine-coded, and does it cover every consequential action the agent can take?
2. Is the policy correct, complete and consistent?
3. Does the gate enforce the policy on every action, before execution?
4. Is the action executed the action allowed, against the state on which authority was evaluated?
5. What residue is acceptable, and who decides?

The paper does not answer the fifth question. As in the earlier papers in this series, that is for each institution, and its regulator, to set.

Identity keeps its place as a necessary foundation. It binds the agent to a principal, which is what lets the gate apply the right authority and lets accountability flow to the right party. Evidence of authorisation keeps its place too. A signed mandate is an input to the policy. Neither is a substitute for the gate.

For banks now blocking agents, the gate offers a route back to transactions. Read-only access is the narrowest such gate. A gate with a fuller policy allows transactions within authority and denies the rest. For supervisors, the gate gives something to assess: a policy that can be inspected, and an enforcement point that can be tested. That is a more tractable object of supervision than the conduct of an agent.

10 Conclusion

For a human, proof of identity bears on behaviour in part through accountability, which influences conduct without guaranteeing it. For an agent, identification does not supply that path. The agent's judgement is a distribution that no consequence attached to its identity bounds, conditioned on context that an adversary may control. Proof of identity is not proof of behaviour.

The inference from strong identity foundations to agentic action within mandate is therefore invalid. Indirect prompt injection gives a case in which every premise holds and the conclusion fails.

Blocking, identification and evidence of authorisation each fall short of a guarantee on conduct.

The guarantee has to sit on the action. An action gate applies machine-coded policy to every recommendation before it becomes an action. Under the conditions in section 6.5, every executed action is then within authority, whatever the agent recommends. The named residue is the policy's correctness, the gate's enforcement and complete mediation, each inspectable and checkable. They are members of the trust residue, which also holds the boundary assumptions set out in section 6.5, including the integrity of the records of authority. Before institutions allow agents consequential autonomy, this is how they can replace one part of the trust that consequence used to support: that what executes stays within authority.

References

[1] Son, H. (2026), 'JPMorgan blocks outside AI agents like Muse, Instinct from account access', CNBC, 13 October 2026. https://www.cnbc.com/2026/10/13/jpmorgan-ai-agents-account-access.html

[2] Wurster, R. (2026), Charles Schwab fall business update, 15 October 2026, transcript. https://content.schwab.com/web/retail/public/about-schwab/schwab_fall_business_update_transcript_10152026.pdf. Interview with Bloomberg Television, 16 October 2026, reported by The Verge. https://www.theverge.com/ai-artificial-intelligence/1013303/charles-schwab-is-blocking-ai-agents-too

[3] Axios (2026), 'Scoop: Unvetted AI agents are accessing U.S. bank accounts', 22 October 2026. https://www.axios.com/2026/10/22/ai-agents-banks-unauthorized-accounts-access. Reuters, 'US banks are blocking third-party AI agents access to customer accounts, sources say', 22 October 2026. https://www.channelnewsasia.com/business/us-banks-are-blocking-third-party-ai-agents-access-customer-accounts-sources-say-6482066

[4] Saunders, T. (2026), 'Banks ban AI from accounts over fears of rogue agents', *The Telegraph*, 8 November 2026. https://www.telegraph.co.uk/business/2026/11/08/banks-ban-ai-from-accounts-over-fears-of-rogue-agents/

[5] Trulioo and PayOS, 'Know Your Agent (KYA): An Identity Framework for Agentic Commerce'. https://www.trulioo.com/resources/white-papers/know-your-agent-an-identity-framework-for-trusted-agentic-commerce

[6] Ant International, Mastercard and Visa (2026), 'Ant International, Mastercard and Visa Initiate Collaboration on Know-Your-Agent Interoperability to Scale Agentic Commerce', press release, Business Wire, 9 September 2026. https://www.businesswire.com/news/home/20260909003891/en/

[7] Anderson, J. P. (1972), *Computer Security Technology Planning Study*, ESD-TR-73-51, Vols I and II, Electronic Systems Division, US Air Force, October 1972 (DTIC AD-758 206 and AD-772 806).

[8] Betts, R. (2026), 'Trust and Proof', 7 September 2026. https://rileybetts.ai/papers/trust-and-proof. DOI: https://doi.org/10.5281/zenodo.24293604. Proof relative to a boundary; local discharge and the trust residue $R(A)$ §6; Trust Displacement Result §6; trust-dominance §7.

[9] Betts, R., Riley, M. and Riley-Betts, J. (2026), 'Goal-Relative Assurance of Dynamically Generated Execution Traces', 11 September 2026. https://rileybetts.ai/papers/goal-relative-assurance. DOI: https://doi.org/10.5281/zenodo.24293776. Floor fixed outside the execution §9; residue migration §21; Goal Discharge Condition §25.

[10] Betts, R., Riley, M. and Riley-Betts, J. (2026), 'Agentic Process Governance', 5 October 2026. https://rileybetts.ai/papers/agentic-process-governance. DOI: https://doi.org/10.5281/zenodo.24293999. Allow is not the claim, and evidence stays separate from the gate's decision §3; the policy machine §5.

[11] Monetary Authority of Singapore (2026), *Safeguards for Agentic Finance at Runtime* (SAFR), white paper, version 1.1, October 2026; version 1.0 published 3 July 2026. https://www.mas.gov.sg/-/media/mas-media-library/development/fintech/ai-safr/safr-v1-1.pdf. Summary: Rajah & Tann. https://www.rajahtannasia.com/viewpoints/mas-publishes-white-paper-for-fis-on-developing-safeguards-for-ai-agents-in-finance/

[12] Entrust (2026), 'Who Authorized This Payment? Trust and Banking in the Age of AI Agents', The Identity Thread podcast, with Mayank Somaiya (Backbase) and Andy Cease (Entrust), 22 October 2026. https://cybersecurity-institute.podcasts.entrust.com/e/who-authorized-this-payment-trust-and-banking-in-the-age-of-ai-agents/

[13] Wells Fargo (2026), 'We Are Updating Our Online Access Agreement', effective 24 November 2026, section 15. https://www.wellsfargo.com/online-banking/updates/. The agreement text: https://www.wellsfargo.com/online-banking/online-access-agreement/upcoming/. Reported by American Banker, 2 November 2026. https://www.americanbanker.com/news/wells-fargo-wants-customers-to-prove-theyve-revoked-ai-agents

[14] FIDO Alliance (2026), 'FIDO Alliance to develop standards for trusted AI agent interactions', 28 April 2026. https://fidoalliance.org/fido-alliance-to-develop-standards-for-trusted-ai-agent-interactions/

[15] South, T., for the OpenID Foundation (2025), 'Identity Management for Agentic AI', sections 2.10, 2.12, 3.4 and 3.5. Published 7 October 2025. https://arxiv.org/abs/2510.25819

[16] South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C. D., Greenwood, D., Chan, A. and Pentland, A. (2025), 'Authenticated Delegation and Authorized AI Agents', arXiv:2501.09674. https://arxiv.org/abs/2501.09674. Position version in *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267, pp. 82211–82231.

[17] Debenedetti, E., Shumailov, I., Fan, T., Hayes, J., Carlini, N., Fabian, D., Kern, C., Shi, C., Terzis, A. and Tramèr, F. (2025), 'Defeating Prompt Injections by Design', arXiv:2503.18813. https://arxiv.org/abs/2503.18813

[18] Shi, T., He, J., Wang, Z., Li, H., Wu, L., Guo, W. and Song, D. (2025), 'Progent: Securing AI Agents with Privilege Control', arXiv:2504.11703. https://arxiv.org/abs/2504.11703

[19] The Money Laundering, Terrorist Financing and Transfer of Funds (Information on the Payer) Regulations 2017, regulation 28. https://www.legislation.gov.uk/uksi/2017/692/regulation/28

[20] Bishop, M., Engle, S., Peisert, S., Whalen, S. and Gates, C. (2009), 'Case Studies of an Insider Framework', *Proceedings of the 42nd Hawaii International Conference on System Sciences*. https://web.cs.dal.ca/~gates/papers/hicss09.pdf. DOI 10.1109/HICSS.2009.104. The definition originates in Bishop, M. (2005), New Security Paradigms Workshop, pp. 75–76.

[21] Brunner, T., Liu, Y.-H. and Pande, M. (2026), 'AI threats in the wild: the current state of prompt injections on the web', Google, 23 April 2026. https://blog.google/security/prompt-injections-web/

[22] Lagerquist, J. (2026), 'BMO, CIBC, RBC, TD Bank block personal AI agents, citing safety and security', CBC News, 29 October 2026. https://www.cbc.ca/news/business/canadian-banks-block-ai-9.7400777

[23] Reuters (2026), 'UBS says does not permit use of external AI models for client transactions', 27 October 2026. https://www.channelnewsasia.com/business/ubs-says-does-not-permit-use-external-ai-models-client-transactions-6495841