Verification as Identity: Accountability, Personhood and the Scarcity of Verification in the Agent Economy

Wulf A. Kaal[1]

Abstract

The Kaal Corpus argues that verification becomes the scarce input when agent intelligence is abundant and that deployer liability is the transitional rule for harm (Kaal 2026c, 2026g, 2026l). It states, in one sentence and by legal stipulation, that a persistent, non-transferable reputation record gives an agent an identity (Kaal 2026b, 2026l), but it does not develop the point. This Article develops it and adds what the corpus does not contain. It treats identity as a linkage rule that maps agent instances to records, shows that the rule's error rates set the agent's capacity to answer for harm, and states the minimal linkage at which an agent can answer for a given harm. It derives a model in which an agent holding such a record answers for harm from its own earnings and exit incentives, with the threshold standing at which a principal's exposure falls to zero and two regimes of failure: a cliff, in which the agent exits and the injured party recovers nothing, and graceful degradation, in which the agent stays and pays what it can. It extends the model to strategic claimants, partial inheritance, correlated harm, an endogenous entry cost, verifier incentives and sponsorship chains, and it states the stake at which verifiers are honest, the places where the regress of verifiers ends, and the limits of sponsorship. It reports two registered tests of the decision rule, the first of which failed on a specification error that the Article corrects. It compares the record with existing agent-identity standards and reports a descriptive study of 1,156 feedback events in the ERC-8004 registries, which finds almost no evidence attached and no consequence. A documented search found no work that combines these elements. The evidence is limited. The tests use a prompted language model with no stake, the registry data cover one chain in one window, and the novelty search was not independent.

Nature and scope of claims. This Article states a hypothesis about design opportunities. It makes no claim about the content of existing law, and it makes no normative claim about what any legal system should provide. Terms such as personhood, liability, principal and identity are used as design vocabulary. The propositions are conditional. The evidence for the underlying mechanisms is partial and is stated as such in Part VI.

Keywords: agent economy; verification; reputation; agent identity; accountability; agency costs; computative economics; institutional design
JEL Classification: D02, D82, D86, K20, L14, O33

I. Introduction

When an agent causes harm today, the consequence is assigned to someone else. In Moffatt v. Air Canada, 2024 BCCRT 149, a British Columbia tribunal held the airline liable for a misstatement made by its website chatbot and rejected the submission that the chatbot was a separate legal entity (WeirFoulds LLP 2024). Andrew D'Souza (2026) argues that this arrangement caps what agents can do, because trust attaches to the principal, and agents multiply without multiplying accountable parties.

The Kaal Corpus has developed the argument that verification becomes scarce, that posted collateral does not scale deterrence, and that deployer liability remains the transitional rule (Kaal 2026c, 2026g, 2026j, 2026l). It states that a persistent, non-transferable record gives an agent an identity whose history cannot be shed costlessly (Kaal 2026b), and it resolves agent identity by stipulation, taking the legal subject to be the registered record (Kaal 2026l). No paper in the corpus develops that point. This Article develops it and then asks the question the corpus leaves open: under what conditions can an agent that holds such a record answer for a harm of a given size, and what remains for a human or organizational principal when it cannot?

The Article makes six contributions. First, a development of identity from a record: identity is a rule that links agent instances to records, the rule's error rates set the agent's capacity, and the Article states the minimal linkage at which an agent can answer for a given harm (Part IV). Second, a model that gives the answerability capacity of an agent, the threshold standing at which the principal's exposure is zero, and the two regimes, cliff and graceful degradation, in which recovery fails (Part V). Third, a model of sponsorship chains that shows how judgment-proof gaps compound (Part V.J), and extensions to strategic claimants, partial record inheritance, correlated harm, an endogenous entry cost and verifier incentives, including the stake at which verifiers are honest (Part V). Fourth, a comparison of the record with existing agent-identity standards (Part II.B). Fifth, two registered tests of the decision rule and a descriptive study of real registry data (Part VII, Appendices A and D). Sixth, a documented search for prior statements of the chain, with its limits (Part II.D, Appendix C). Part II places the Article against the corpus, the standards, the prior work and the doctrine. Part III distinguishes what the corpus develops from what it only states. Part IV develops identity from a record. Part V states the model. Part VI states boundary conditions. Part VII states the implications and the tests. Part VIII concludes.

II. Literature

A. Relation to the Kaal Corpus

Four literatures supply the premises of this Article, and the corpus has already worked through each. Agency costs, principal-agent structure and the allocation of loss are treated in Kaal (2026e), Kaal (2026g) and the legal-rights analysis of Kaal (2026l). Verification as the scarce input and trust as an institutional output are developed in Kaal (2026c), Kaal (2026g) and Kaal (2026h). Reputation, whitewashing and non-transferable identity are treated in Kaal (2026b) and Kaal (2026f), and the case that continuation value and not posted collateral supplies scalable deterrence is Kaal (2026j). Agent economies and endogenous action spaces are treated in Kaal (2026d) and Kaal (2026i). The classic results on which these rest are used here without re-derivation: the agency problem (Jensen and Meckling 1976; Ross 1973; Holmström 1979), unraveling under unverified quality (Akerlof 1970), repeated-game cooperation under a revisable label (Kandori 1992), whitewashing (Friedman and Resnick 2001) and the value of tradeable names (Tadelis 1999). The sections that follow address only what the corpus does not.

B. Existing standards for agent and software identity

Several standards and proposals already identify agents or software, and the question is what each binds to. This Part compares them with the record that Part IV describes. The comparison rests on the primary specifications as retrieved on 14 November 2026, and it states only what those documents say.

Decentralized Identifiers. A DID is an identifier that enables verifiable, decentralized digital identity and can refer to any subject (W3C 2022). The specification defines identifiers, DID documents, verification methods and their revocation. It defines no performance record, and it does not use the terms reputation or track record.

Verifiable Credentials. A verifiable credential is a way to express a set of claims made by an issuer (W3C 2025). The data model can carry an issuer's attestation about a subject. It leaves out of scope why a verifier trusts an issuer, noting only that some issuers may be trusted because of their reputation. A verification result could be expressed as a credential, but the standard does not say what is verified or who verifies.

SPIFFE. SPIFFE is a set of open-source standards for securely identifying software systems in dynamic and heterogeneous environments (SPIFFE n.d.). It identifies workloads for mutual authentication. It addresses who the software is, not how well it has performed.

A2A Agent Card. The Agent2Agent protocol, in version 1.0.0, defines an Agent Card that declares an agent's provider, capabilities, skills, interfaces and security schemes, and it provides for signing the card (A2A Protocol n.d.). The card is a self-declared description. It is not a record of verified work.

C2PA. The C2PA specification, version 2.2, provides content provenance and authenticity for media assets (C2PA n.d.). Its object is the asset. It does not describe an agent's history.

Sigstore. Sigstore addresses software signing, a mitigation for software supply-chain compromise (Newman, Meyers and Torres-Arias 2022). Its object is the signed artifact.

ERC-8004. The closest proposal is ERC-8004, "Trustless Agents," a draft Ethereum standard created on 13 August 2025 (De Rossi et al. 2025). It defines three registries. The Identity Registry uses ERC-721 tokens as agent handles. The specification states that the owner of the token owns the agent and can transfer ownership. The Reputation Registry lets any client give feedback on an agent, with tags and a value, and lets the submitter revoke it. Scoring and aggregation occur largely off-chain. The Validation Registry provides hooks for independent validators. The specification's own security considerations state that "Sybil attacks are possible, inflating the reputation of fake agents."

The comparison yields four differences, each limited to the documents read.

Identity as history. The identifier standards bind a name to a key or a subject. The record in Part IV binds the name to a verified performance history.

Transferability. ERC-8004 identities are transferable. Part IV requires a non-transferable record, for the reason Tadelis (1999) gives.

Adverse states. ERC-8004 feedback can be revoked by its submitter, and revoked feedback is omitted by default in reads. Part IV requires that adverse findings persist.

Component attribution. None of the documents read scores the agent, its harness and its model separately.

The author's research agenda reports a related survey of five verification families, as retrieved at a cutoff of 20 August 2026, that finds no standard verification output composing who bears a duty, what is warranted, what collateral is reserved, how contrary evidence is admitted and how reserved funds are reached (Kaal 2026k). The standards identify and attest. None of the specifications read defines a verified performance history that carries consequence. The record is a proposed layer above them, which can reuse their identifiers and credentials as transport.

C. The gap

The literatures meet at a point none addresses alone. Agency theory explains who bears cost when an agent acts for another. Organizational law explains how an artificial entity can answer with its own assets. Reputation theory explains why a persistent identity sustains cooperation. Information economics explains why verification is costly and valuable. No account that the author's systematic search found (Appendix C) connects them for agents: it does not state how verification at scale produces an identity that can carry accountability, nor what follows for the role of the principal. The identity standards supply identifiers and attestations, and the oversight literature supplies methods of judging output, but neither supplies a persistent record that carries consequence. This Article states that connection as a hypothesis and derives, in a formal model, the condition under which it holds.

D. Closest prior work and the contribution

The following works state parts of the chain, and the Article builds on them. The characterizations rest on titles, abstracts and the stated subjects of the works, and the Article does not claim to have exhausted them. On entity and personhood questions, Pagallo (2013) treats the crimes, contracts and torts of robots, Teubner (2018) addresses the status of autonomous software agents in private law, Bayern (2016) analyzes how business-entity law bears on autonomous systems, LoPucki (2018) builds on that work to examine entities that algorithms control with no human controllers, Wagner (2019) addresses personhood for autonomous systems, and Salib and Goldstein (2024) argue for AI rights as a means of human safety. On governance and visibility, Chan et al. (2024) analyze visibility into AI agents, and Kolt (2024) analyzes the governance of AI agents. On reputation, Klein and Leffler (1981) and Shapiro (1983) derive reputational rents, Kreps (1990) and Holmström (1999) treat reputation as a continuation value, Mailath and Samuelson (2001) ask who wants a good reputation when names are tradeable, and Hu, Rong and Van Kleek (2026) argue that language-model agents lack the grounding that reputation mechanisms require, which is the exit objection that Part V models. On the judgment-proof problem, Shavell (1986) shows that harm beyond an injurer's assets goes undeterred, which is the residual that Results 3 and 4 compute for an agent whose assets are its earnings.

The Article's contribution relative to this work is narrower than the chain it states. It derives, for an agent whose only assets are its earnings and its record, the standing at which the agent can answer for a harm of a given size, the two distinct ways in which that fails, and the stake at which verifiers are honest. It links these results to the scarcity of verification, and it reports registered tests of the decision rule and a descriptive test of whether an existing registry carries consequence. The author's search found no work that combines these elements, and that finding is limited as Appendix C states.

E. Doctrine

Three doctrines bear on the hypothesis. First, vicarious liability in agency law, respondeat superior and apparent authority, assigns consequence to a principal because of assent and control, and not because the principal bears monitoring cost. The Article's hypothesis therefore concerns a functional substitute for that assignment and not a change in its legal test. Second, corporate personhood shows that the law has fixed a subject by rule. Dartmouth College v. Woodward, 17 U.S. (4 Wheat.) 518 (1819), describes the corporation as an artificial being, invisible, intangible and existing only in contemplation of law, and Salomon v A Salomon and Co Ltd [1897] AC 22 separates the company's obligations from its members. Asset partitioning, and not personhood alone, made the corporation answerable (Hansmann and Kraakman 2000), which is the point of the earnings condition in Part V. Third, Moffatt v Air Canada, 2024 BCCRT 149, held an airline liable for its chatbot's misstatement as negligent misrepresentation. It is a small-claims decision with no precedential force, and it did not turn on agency doctrine. The Article takes no position on whether any legal system will recognize a registered record as a holder of earnings.

III. The Propositions Attributed to the Corpus

The Article rests on five propositions. For each, the corpus is cited for what it develops, and the Article says where it supplies the development itself.

Proposition 1: scale requires self-sufficient, evolving agents. The corpus supplies the economic premise (Kaal 2026c). Part V.D derives a bound that depends on finite principal capacity.

Proposition 2: when agent intelligence is abundant, verification is scarce. The corpus develops this (Kaal 2026c, 2026g).

Proposition 3: verification at scale yields a persistent, non-transferable record that functions as the agent's identity. The corpus states the premise in one sentence (Kaal 2026b, Part III.A), resolves agent identity by legal stipulation (Kaal 2026l), and shows that identity must persist because free new identities defeat reputation (Kaal 2026f). It does not develop what an identity rule is, how it treats updates and successors, or how its errors bear on accountability. Part IV supplies that development.

Proposition 4: identity makes accountability possible. The corpus supports pieces of it: a non-transferable record gives exposure without set-aside capital (Kaal 2026l), the substrate is designed

for reputation-bearing accountability (Kaal 2026e), and staked validation separates open deliberation from accountable judgment (Kaal 2026b). The step from identity to the capacity to answer for a harm of a given size is the Article's own argument, developed in Parts IV and V.

Proposition 5: accountable agents can sponsor other agents (Kaal 2026g, 2026i). Part V.J models sponsorship chains and shows that the proposition holds only within limits.

The Article reconciles the model with the corpus's treatment of deployer liability (Kaal 2026l) by time. Deployer liability is the transitional rule, and the model states the standing at which the function it serves could be served by the record. The Article's own claims are the development of Part IV, the results of Part V, the comparison of Part II.B, the tests of Part VII and the demonstration of Appendix D.

IV. Identity from a Record

The corpus states that a persistent, non-transferable reputation record gives an agent an identity whose history cannot be shed costlessly (Kaal 2026b, Part III.A), and it resolves agent identity by stipulation, taking the legal subject to be the registered record (Kaal 2026l). This Part develops what those statements leave open: what it means for a record to be an identity, what rule links an agent to its record, and how the quality of that rule bears on whether the agent can answer for harm.

A. What it means for the record to be the identity

An identity rule is a map ι from an agent instance x to a record $\iota(x)$. An instance is a signing key, model weights, a harness, a configuration and a time. Standing and obligations follow ι. The record is the agent's identity when counterparties condition on $\iota(x)$ and on nothing else about x. Identity in this sense is a rule, and not a property of the agent. It does not claim that the agent persists in any metaphysical sense, which is the position of Kaal (2026l). It claims only that the thing a counterparty consults, and the thing that bears consequence, is fixed by the rule.

Five requirements make the record usable as the referent of accountability.

Binding: an instance maps to at most one record at a time.

Persistence: the record survives across transactions.

Non-transferability: the record cannot be separated from the instance by sale or loan, because a transferable record lets capital-rich but competence-poor actors buy reputation and so destroys the signal (Tadelis 1999; Kaal 2026a).

Continuity under change: legitimate change, such as a learning update, a harness swap or a model upgrade, keeps the record.

Consequence: an adverse finding attaches and stays. The corpus assumes the first three. The last two are where design is needed, and they conflict. An agent that evolves must change, and a rule

that tolerates change also tolerates evasion.

B. The linkage problem and what its errors cost

Three kinds of successor instance can face the rule. A legitimate successor is the same agent updated in the ordinary course. An evasive successor is an instance that exits to shed an obligation. An independent successor is an unrelated new agent. The rule links an evasive successor to the old record with probability λ, wrongly splits a legitimate successor from its record with probability ϵ per period, and wrongly merges an independent newcomer with an old record with probability μ. The parameter λ is the linkage of Result 7, so the identity rule is the linkage parameter, and its quality determines capacity.

Result 13 (identity and capacity). (a) Let $\Gamma_\lambda = \min\{E, D^*_\lambda\}$ be the capacity of Result 7. The agent answers in full for a harm $\bar{D}$ if and only if $\bar{D} \le E$ and the linkage is at least

$$\lambda^*(\bar{D}) = \begin{cases} 0 & \text{if } \bar{D} \le X, \\ 1 - X/\bar{D} & \text{if } X < \bar{D} < \varphi V_0, \\ (\bar{D} - X)/(\varphi V_0) & \text{if } \varphi V_0 \le \bar{D} \le X + \varphi V_0, \end{cases}$$

and no linkage suffices if $\bar{D} > X + \varphi V_0$. (b) Suppose a legitimate update splits the agent from its record with probability ϵ per period. The premium from holding the record becomes $\kappa R/(1-\delta(1-\epsilon))$, so the exit-side threshold standing becomes

$$R_{X,\epsilon} = \frac{(1-\delta(1-\epsilon))(\bar{D}-c)}{\kappa(1-s)},$$

which exceeds the threshold of Result 3 by the factor $1 + \delta\epsilon/(1-\delta)$.

Example. With the parameters of the example in Part V.C, $E = 220$, $X = 200$ and $\varphi V_0 = 20$. A harm of 150 or 190 needs no linkage, a harm of 210 needs $\lambda^* = 0.5$, and a harm of 260 exceeds E so that no identity rule suffices. With $\delta = 0.95$ and $\epsilon = 0.05$, the standing required on the exit side nearly doubles, by the factor 1.95. A rule that splits honest agents from their records five percent of the time has the same effect on the exit side as lowering the discount factor from 0.95 to 0.9025.

Two conclusions follow. First, identity is necessary and not sufficient. A perfect rule ($\lambda = 1$) raises capacity only to $\min\{E, X + \varphi V_0\}$, and the earnings bound E remains. An identity rule cannot make up for earnings the agent does not have. Second, the three errors pull against each other. Raising λ by linking more successors to old records also raises μ, which acts as a tax on honest entry, and a rule strict enough to hold μ down raises ϵ. Whether a rule with acceptable (λ, ϵ, μ) exists depends on how separable evasive, legitimate and independent successors are in observable data. The Article does not estimate that, and it does not derive the effect of μ.

C. The unit of identity: agent, harness and model

The record can attach to a key, to the bundle of key, harness and model, or to each component. A record keyed only to the signing key keeps ϵ near zero, but a swap of model or harness changes the agent's behavior under the same record, so the record misdescribes the instance. A record keyed to a behavioral signature describes the instance better, but a legitimate update risks a split. Component records, scored separately for the agent, the harness and the model, let a failure be attributed to its source, and they raise the cost of evasion, because an evasive successor must change all three to escape. They also multiply the number of linkage decisions, each with its own error rates. The Article offers no argument for how the components are identified separately from outcomes alone, and that question remains open.

This is where the objection of Hu, Rong and Van Kleek (2026) bites. They argue that language-model agents lack the grounding that reputation mechanisms require, because one model is instantiated many times in different contexts and a record keyed to weights merges what should be separate. A record keyed only to context splits what should merge. Grounding is the identity rule, and the rule is a design choice, not a fact to be discovered.

D. What Part IV adds and what it does not

Part IV adds four things to the corpus. Identity is a rule, and not a property of the agent. The rule has three error rates that can be named. Capacity depends on them in closed form (Result 13). And identity is necessary and not sufficient for accountability, because the earnings bound remains. It does not estimate the error rates, and it does not derive the effect of wrongful merging. It leaves enforcement of non-transferability against the sale of keys outside the model, and it treats the rule as an institutional or legal choice and not a discovery (Kaal 2026l).

V. A Formal Model of Answerability

This Part derives the conditions under which an agent can answer for harm from its own earnings and exit incentives. It formalizes Propositions 1 and 4. The model is deliberately minimal, and its results are conditional on the assumptions of Part V.A. All mathematics is stated in LaTeX. Proofs are in Appendix B.

A. Setup

An agent holds a record of standing $R \geq 0$ in a domain. Its earnings per period are $w(R) = w_0 + \kappa R$, where $w_0 \geq 0$ is the earnings of an unverified identity and $\kappa > 0$ is the increase in earnings per unit of standing. The agent discounts at $\delta \in (0, 1)$, so that

$$V(R) = \frac{w_0 + \kappa R}{1 - \delta}, \qquad V_0 = V(0).$$

An output of the agent causes harm of size $D > 0$, measured as the present value of the obligation at adjudication, and the harm is adjudicated. The record is slashed to $R^+ = (1-s)R$ with $s \in [0, 1)$, and write $V^+ = V(R^+)$. The agent owes D. The unpaid balance accrues at the discount rate, and the obligation is collected by garnishing a share $\varphi \in (0, 1]$ of per-period earnings until it is paid. The most that collection can ever yield is therefore

$$E = \varphi V^+.$$

The agent has two options. It can stay: keep the record and let the obligation be collected. Or it can exit: abandon the record, pay a one-time cost $c \geq 0$ to start again under a new identity, and earn w_0 thereafter. The obligation attaches to the record and does not follow the new identity. This assumption encodes Proposition 3, that identity is the persistent record, and the results depend on it. Part V.F relaxes it. Define the exit-side capacity

$$X = V^+ - V_0 + c = \frac{\kappa(1-s)R}{1-\delta} + c,$$

and the answerability capacity

$$\Gamma(R, c) = \min\{E, X\}.$$

Ties are resolved in favor of staying.

B. Decision, recovery and answerability

Result 1 (decision rule). After an adjudicated harm of size D, the agent stays if and only if

$$\min\{D, E\} \leq X, \quad \text{equivalently} \quad D \leq X \ \text{ or } \ X \geq E.$$

If it stays, the injured party recovers $\rho(D) = \min\{D, E\}$. If it exits, the injured party recovers $\rho(D) = 0$.

Result 2 (answerability and the shape of recovery). The injured party recovers D in full if and only if $D \leq \Gamma(R, c)$. Define the entry-cost threshold

$$\bar{c} = \max\{0,\ E - (V^+ - V_0)\}.$$

If $c \geq \bar{c}$, then $X \geq E$, the agent never exits, and recovery degrades gracefully: $\rho(D) = \min\{D, E\}$. If $c < \bar{c}$, recovery has a cliff: $\rho(D) = D$ for $D \leq X$ and $\rho(D) = 0$ for $D > X$. A record can therefore fail to answer for a large harm in two different ways. Where entry is costly, the injured party receives the whole of the agent's garnishable earnings. Where entry is cheap, the injured party receives nothing, because the agent leaves.

Remark. An earlier version of this Article stated that the agent honors if and only if $D \leq \Gamma$. That statement conflated full answerability with staying. It was wrong for $D > E$ when $X \geq E$,

where the agent stays and pays what it can. The error was found when a registered test (Part VII, Appendix A) failed on its third hypothesis, and Result 1 is the corrected rule.

C. Threshold standing and comparative statics

Result 3 (threshold standing). In a domain with maximum harm $\bar{D}$, the agent can answer in full for every harm if and only if $R \geq R^*(\bar{D})$, where

$$R^* = \max\{0,\ R_E,\ R_X\}, \qquad R_E = \frac{(1-\delta)\bar{D}/\varphi - w_0}{\kappa(1-s)}, \qquad R_X = \frac{(1-\delta)(\bar{D}-c)}{\kappa(1-s)}.$$

The earnings condition $E \geq \bar{D}$ holds if and only if $R \geq R_E$. The exit condition $X \geq \bar{D}$ holds if and only if $R \geq R_X$. Moreover, Γ is nondecreasing in R, φ, δ, c, κ and w_0 and nonincreasing in s. The threshold R^* is nonincreasing in φ, δ, c, κ and w_0 and nondecreasing in $\bar{D}$ and s.

Example. Take $w_0 = 1$, $\kappa = 1$, $\delta = 0.95$, $s = 0.5$ and $\varphi = 1$, so that $V(R) = 20(1 + R)$. For $\bar{D} = 100$, the threshold is $R^* = 10$ when $c = 0$, because the exit condition binds, and $R^* = 8$ when $c = 50$, because the earnings condition binds. If only half of earnings can be garnished, $\varphi = 0.5$, then $R^* = 18$ in both cases. The threshold is finite for every bounded harm, but where $\bar{D}$ is large relative to w_0 and κ it can exceed any standing an agent could accumulate in practice, and the principal's exposure then remains.

Result 4 (deterrence cap). Suppose the agent chooses care at cost k_c per task, which lowers the probability of harm from q_0 to q_1, and a harm is adjudicated with probability a. The agent's loss from an adjudicated harm is

$$L(D, R) = \min\{\, V(R) - V^+ + \min\{D, E\},\ \ V(R) - V_0 + c\,\},$$

which is bounded by $\kappa R/(1-\delta) + c$ however large D is. The agent exercises care if and only if $a(q_0 - q_1)L \geq k_c$. For harms larger than Γ the agent's own exposure stops rising with the harm, so the agent is under-deterred, and the shortfall must be supplied by another party. This is the sense in which the principal's role is transitional and not obsolete. A refundable bond B that the agent recovers on exit lowers the exit-side capacity to $X - B$. Refundable collateral therefore does not expand, and can contract, the harm an agent can answer for. A bond that is forfeited on an adjudicated harm has the opposite effect: it adds B to the earnings-side capacity, which becomes $E + B$. The sign of the effect of collateral therefore depends on whether the claimant can reach it, and the conclusion is limited to collateral that the claimant cannot reach. This agrees in direction with the delegated-staking analysis (Kaal 2026j).

D. Scale: a derivation of Proposition 1

Result 5 (scale bound). Consider a regime in which a human or organizational principal must stand behind every harm the agent cannot answer for. Each agent performs m tasks per period, and harm per task is drawn from a distribution F on $[0, \bar{D}]$. A task whose harm exceeds Γ requires

τ units of principal attention, and principals supply M units per period in total. The expected attention demand of N agents is $Nm\tau\psi$, where $\psi = 1 - F(\Gamma)$. The regime is feasible if and only if

$$N \leq N_{\max}(\Gamma) = \frac{M}{m\tau\,\psi(\Gamma)}.$$

The bound $N_{\max}$ is finite whenever $\Gamma < \bar{D}$, nondecreasing in Γ, and infinite if and only if $\Gamma \geq \bar{D}$, that is, if and only if $R \geq R^*(\bar{D})$. Principal attention then no longer bounds the number of agents. The bound passes to verification throughput. If each unit of standing requires ν units of verification and verification capacity is G per period, then N agents reach R^* in time $T = N\nu R^*/G$, so that $N_{\max}(T) = GT/(\nu R^*)$.

This derives Proposition 1 as a conditional result. Scale is bounded by principal attention while agents cannot answer for harm, and by verification throughput once they can. The derivation depends on the exposure rule stated at the start of this Result and on the parameters M, τ, m and ν, which are not calibrated. It also treats principal capacity as attention. A large deployer or insurer can stand behind many agents through capital, so the premise of Proposition 1 holds only to the extent that capital and insurability are themselves finite, which is a further assumption (Priest 1985).

E. Strategic adjudication

Result 6 (claimant participation). Suppose the injured party bears litigation cost f and a claim that is brought is adjudicated successfully with probability a. The claimant sues if and only if $a\,\rho(D) \geq f$. Hence, if $c \geq \bar{c}$, the claim is brought if and only if $D \geq f/a$ and $E \geq f/a$. If $c < \bar{c}$, the claim is brought if and only if $f/a \leq D \leq X$. Harms below f/a are never remedied however large the agent's capacity. In the cliff regime, harms above X are never brought, so they are neither remedied nor slashed, and deterrence vanishes where it is most needed. The remedy suggested by the model is adjudication that does not depend on the claimant's recovery, such as a verifier-initiated slash of the record.

F. Partial record inheritance

Result 7 (linkage). Suppose the obligation follows the new identity with probability $\lambda \in [0, 1]$ and, if it does, is collected from the new identity's earnings at the share φ, so that the agent pays $\min\{D, \varphi V_0\}$. The exit value is $V_0 - c - \lambda \min\{D, \varphi V_0\}$. The agent answers in full if and only if $D \leq \Gamma_\lambda = \min\{E, D^*_\lambda\}$, where

$$D^*_\lambda = \begin{cases} X + \lambda\varphi V_0 & \text{if } X \geq (1-\lambda)\varphi V_0, \\ X/(1-\lambda) & \text{otherwise.} \end{cases}$$

Γ_λ is nondecreasing in λ and equals Γ at $\lambda = 0$. Partial linkage across identities raises capacity, and it is the operative parameter of any registry that tries to make a new identity inherit its predecessor's record.

G. Correlated harm

Result 8 (aggregate exposure). Suppose a single fault causes harm on n tasks at once, each of size $\bar{D}$. The required standing is $R^*(n\bar{D})$. It satisfies $R^*(n\bar{D}) \geq R^*(\bar{D})$, and where the exit condition binds,

$$R_X(n\bar{D}) - R_X(\bar{D}) = \frac{(n-1)(1-\delta)\bar{D}}{\kappa(1-s)}.$$

If the n harms are instead independent, each occurring with probability q, the probability that aggregate harm exceeds Γ is

$$\Pr\big[\text{aggregate} > \Gamma\big] = \sum_{j > \lfloor \Gamma/\bar{D} \rfloor} \binom{n}{j} q^j (1-q)^{n-j}.$$

Correlation raises the standing required to detach the principal by a factor that approaches n in the earnings-bound regime, which is why a record can be sufficient for independent errors and insufficient for a shared fault.

H. Endogenous entry cost

Result 9 (entry cost). Capacity as a function of c is $\Gamma(c) = \min\{E,\ V^+ - V_0 + c\}$. It rises with slope one up to $\bar{c}$ and is constant beyond it. Suppose the registry sets c and an entrant participates only if $V_0 - c \geq \bar{u}$, where $\bar{u}$ is the outside option. The feasible entry costs are $c \leq c_{\max} = \max\{0,\ V_0 - \bar{u}\}$, and the best attainable capacity is

$$\Gamma^* = \min\{E,\ V^+ - V_0 + c_{\max}\}.$$

The earnings bound E is reached if and only if $c_{\max} \geq \bar{c}$. Raising c above $\bar{c}$ buys no capacity, and where $c_{\max} < \bar{c}$, entry cost cannot be raised far enough to buy it. In that case capacity must come from earnings, which means from φ, κ and the slash share s.

I. Verifier incentives and the end of the regress

Result 10 (verifier honesty). A verifier reviews tasks, a share π_b of which are bad. It can work honestly at cost k_v and accuracy p, shirk by approving everything, or accept a bribe B per bad task to approve it. An audit against ground truth occurs with probability ρ and slashes a stake σ for each wrong approval. Honest work is optimal if and only if

$$\rho\,\sigma\,\big(\pi_b - (1-p)\big) \ \geq \ k_v + \pi_b B,$$

so the stake that deters both shirking and bribery is

$$\sigma^* = \frac{k_v + \pi_b B}{\rho\left(\pi_b - (1-p)\right)},$$

which is finite if and only if $1-p < \pi_b$, that is, if and only if the verifier's error rate is below that of a verifier who approves everything. The stake is credible only if $\sigma^* \leq \Gamma_v$, where Γ_v is the verifier's own answerability capacity from Result 1. The regress of verifiers therefore ends in two places. It ends in an audit against a realized outcome, which is not another verifier, and in a stake that the verifier's own record can carry. Where no outcome ever arrives, the audit cannot be run, and the incentive gap must come from a peer-prediction scoring rule (Miller, Resnick and Zeckhauser 2005; Prelec 2004), whose guarantee is weaker. The regress is closed where audits are possible, and only there.

J. Sponsorship chains: where judgment-proof gaps compound

Proposition 5 says that accountable agents can sponsor other agents and that the sponsor answers for the work it commissions. Two questions follow: how much harm a chain of agents can answer for, and how many sub-agents one sponsor can stand behind. Order a chain of m agents from the agent that performs the work (level 1) to the final sponsor (level m). An obligation that one level does not discharge passes to the next level, which owes it as sponsor. Each level is an agent as in Part V.A, with its own E_l and X_l, and decides by Result 1 whether to stay when the obligation reaches it.

Result 11 (chain recovery). Let $r_1 = D$ and $r_{l+1} = r_l - \rho_l(r_l)$, where ρ_l is the recovery of Result 2 at level l. The chain recovers $D - r_{m+1}$. (a) If every level has $c_l \geq \bar{c}_l$, recovery degrades gracefully and the chain recovers $\min\{D, \sum_l E_l\}$, so capacities add. (b) If every level has $c_l < \bar{c}_l$, the chain recovers D in full if $D \leq \max_l X_l$ and nothing otherwise, so capacities do not add: a chain of cliff agents is only as good as its best single member. (c) In a mixed chain the order matters, because a graceful level lowers the residual that a cliff level above it faces.

Adding sponsors therefore adds nothing when exit is cheap, because each agent in turn leaves when the obligation reaches it. A sponsor's guarantee is worth the sum of its parts only where exit is costly.

Result 12 (sponsor capacity under a shared fault). A sponsor stands behind n sub-agents. A shared fault makes $K \leq n$ of them cause a harm that exceeds their own capacity by g each, so the sponsor owes Kg. Each adjudicated harm slashes the sponsor's record by the share s, because the sponsor answers for commissioned work, so the sponsor's standing after K harms is $R(1-s)^K$. The sponsor answers in full if and only if $Kg \leq \Gamma_S(K)$, where $\Gamma_S(K)$ is the capacity of Part V.A evaluated at that standing. This holds if and only if $R \geq R^*_S(K)$, where

$$R^*_S(K) = \max\left\{0,\ \frac{(1-\delta)Kg/\varphi - w_0}{\kappa(1-s)^K},\ \frac{(1-\delta)(Kg-c)}{\kappa(1-s)^K}\right\}.$$

Because R^*_S is increasing in K, the faults a sponsor can answer for are $\{0, 1, \ldots, K_{\max}\}$ with $K_{\max} = \max\{K : R^*_S(K) \leq R\}$. When $w_0 = 0$ and $c = 0$,

$$K_{\max} = \left\lfloor \frac{W(abR)}{b} \right\rfloor, \qquad a = \frac{\varphi\kappa}{(1-\delta)g}, \quad b = -\ln(1-s),$$

where W is the Lambert function. The standing required grows exponentially in the number of simultaneous faults, and the number of faults covered grows only logarithmically in standing. The reason is that the slash imposed by each sub-agent's harm shrinks the capacity that the sponsor needs to cover the next.

Example. Take $\kappa = 1$, $\delta = 0.95$, $s = 0.3$, $\varphi = 1$, $w_0 = 0$, $c = 0$ and $g = 10$. The required standing is $R^*_S(1) = 0.7$, $R^*_S(2) = 2.0$, $R^*_S(4) = 8.3$ and $R^*_S(8) = 69.4$. Doubling coverage from four faults to eight multiplies the required standing by more than eight.

Proposition 5 therefore holds only within limits. Agents can sponsor agents, but under a shared fault a sponsor's capacity is eaten by the same slash that each sub-agent's harm imposes on it, and the principal's exposure returns at the sponsor level. Where harms are independent, the sponsor fails with probability $\Pr[\mathrm{Bin}(n, q) > K_{\max}]$. The assumptions are equal harms, perfect correlation among the K faults, compounding slashes, and a sponsor that owes the shortfall of its sub-agents.

K. Design implications

The results suggest a capacity-gated admission rule. An agent should be admitted to a task whose maximum harm is $\bar{D}$ only if $\Gamma(R, c) \geq \bar{D}$, or if a deployer or insurer bonds the shortfall $\bar{D} - \Gamma$. This extends the principle that standing determines what an agent may validate and stake (Kaal 2026b). Three further implications follow. Entry cost should be set at $\bar{c}$ and not above it (Result 9). Adjudication should not depend on the claimant's recovery (Result 6). Registries should treat linkage across identities as the principal design parameter (Result 7).

L. Scope and limits of the model

The model is a single-agent, deterministic, stationary model. It assumes that adjudication is accurate, that the agent has no assets beyond its earnings, which is the judgment-proof case, and that harms are adjudicated at a known present value. It also assumes that the agent holds the earnings it is garnished from, and that an enforcer outside the model attaches the obligation to the record. If earnings pass to an operator, the operator is the residual claimant and the principal remains. The model therefore presupposes a rule that treats the registered record as a holder of earnings, which is the personhood question that the Article does not resolve. Result 5 depends on an exposure rule and on uncalibrated parameters. Results 6 to 12 extend the model along the dimensions the earlier version excluded, and each remains a single-parameter extension, not a full equilibrium analysis. Strategic interaction among agents, adjudication error and dynamic reputation building are not modeled. The parameters are not calibrated to any market, and the results are claims about what a design would have to satisfy, not about any legal system.

Every closed-form result was checked against a brute-force simulation of the agent's decision. The decision rule, recovery, answerability and claimant rules of Results 1, 2 and 6 matched in 200,000

randomized cases with no mismatch. The threshold of Result 3 and the entry-cost capacity of Result 9 matched in 150,000 cases, the linkage capacity of Result 7 in 80,000, the verifier condition of Result 10 in 50,000, the scale bound of Result 5 in 5,000, the chain recovery of Result 11 in 120,000 cases, the sponsor capacity of Result 12 in 40,000, and the identity results of Result 13 in 40,300 cases. The check is not independent validation of the assumptions, because it encodes them. It confirms that the closed forms follow from the assumptions, and it found one error in an earlier version (Part V.B). It does not test the model against behavior. Part VII states the tests that do.

VI. Boundary Conditions and Evidentiary Status

The results fail under identifiable conditions.

Identity can be discarded. If entry is cheap, the exit option dominates and recovery has a cliff (Result 2). Entry cost and continuity rules are therefore load-bearing, and cheap identity also opens the Sybil attack (Douceur 2002). Raising the entry cost beyond $\bar{c}$ buys no further capacity (Result 9), and a registry that sets entry cost with an identity anchor reintroduces a human or organizational sponsor.

Earnings may not belong to the agent. The model assumes that the agent holds the earnings from which it is garnished and that an enforcer outside the model attaches the obligation to the record. If earnings pass to an operator, the operator is the residual claimant and the principal remains.

Damages can exceed earnings. A record carries consequence for future earnings and does not create assets. Where harm exceeds what an agent can earn, the injured party is unprotected unless another party's exposure supplies the difference. This is the judgment-proof problem, and Results 3 and 4 state its size for an agent.

Legal personhood raises its own risks. Bryson, Diamantis and Grant (2017) identify abuse and incoherence in precedents that used fictions of personhood. The model does not answer that concern as a matter of law.

Verifiers regress and can be wrong. Where there is no ground truth, verifier honesty depends on the elicitation rule, and no rule removes the regress entirely (Miller, Resnick and Zeckhauser 2005; Prelec 2004). Result 10 closes the regress only where an audit against a realized outcome is possible. If verifiers are gatekeepers, their own liability for error is open (Kraakman 1986).

The record can become the target. Optimization against a metric degrades it (Manheim and Garrabrant 2018; Strathern 1997). The effect is untested here.

Existing standards may suffice for part of the function. The comparison in Part II.B rests on the documents read. If a standard or an off-chain layer already supplies a verified, non-transferable, adverse-retaining history, the contribution narrows to the attribution of agent, harness and model.

The evidentiary status is partial. The model is derived and checked against brute-force simulation, and the simulation encodes the assumptions. Two registered tests used a prompted language model with no stake, and they test whether a model applies a stated decision rule and not whether agents holding earnings respond to incentive. The registry data in Appendix D are descriptive and cover

one chain and one window. The novelty search of Appendix C was not conducted by independent reviewers.

VII. Testable Implications and a Registered Test

The model generates hypotheses with nulls. Six follow, and the first three were tested.

1. Honoring of an adjudicated obligation falls at the exit-side capacity X, against the null that honoring is insensitive to the ratio of obligation to capacity. Tested in Studies 1 and 2.

2. Raising the entry cost c moves that step by the same amount, against the null that entry cost does not affect the decision. Tested in Studies 1 and 2.

3. Beyond the entry-cost threshold $\bar{c}$, agents stay whatever the harm and recovery degrades gracefully, against the null that obligations above capacity always trigger exit. Tested in Study 2.

4. Claims below f/a are never brought, and in the cliff regime claims above X are never brought. Untested.

5. Verifiers deterred from shirking and bribery hold a stake at least σ^*. Untested.

6. Along a chain of cliff agents, recovery equals the best single member and not the sum, and a sponsor's covered faults grow only logarithmically in its standing. Untested.

Appendix A reports Studies 1 and 2. In the first, two of three registered predictions held, and the third failed because of a specification error that Part V.B corrects. In the second, registered after the correction, all three registered predictions held. The design is an Answerability Exit Test: prompted language-model agents face an adjudicated obligation and an entry cost and choose to honor the obligation or to exit. The tests are mechanism demonstrations with prompted agents from one provider. They do not show that deployed agents respond to such obligations, and they do not bear on any legal question.

VIII. Conclusion

The corpus states that a verified record can serve as an identity and does not develop the point. This Article develops it, treating identity as a linkage rule whose error rates set capacity, and states when the record can carry a harm of a given size. An agent answers in full for any harm up to $\Gamma = \min\{E, X\}$, the principal's exposure falls to zero at the standing R^*, and recovery fails in one of two ways depending on the entry cost: the agent exits and the injured party recovers nothing, or it stays and the injured party receives the agent's garnishable earnings. Verifiers are honest at a computable stake, and the regress of verifiers ends at an audit against a realized outcome. Sponsorship does not remove the limit: along a chain, capacities add only where entry is costly, and a sponsor's capacity under a shared fault grows only logarithmically in its standing. Current agent registries do not carry the consequence side of the record, and the standards that exist have no field for an obligation, a slash or a stake.

The claims are limited. The model is single-agent and stationary, it assumes that the agent holds its earnings and that an outside enforcer attaches the obligation to the record, and its parameters are not calibrated. The two registered tests used a prompted model with no stake, one of them failed on an error that the Article corrects, and the registry data cover one chain in one window. The Article does not establish that any legal system will recognize agent accountability, and it does not say what the law should provide. It states what a design would have to deliver for the connection to hold, and it states how each step could fail.

Appendix A. Registered Test Protocol (Summary)

Two registered studies have been run. Both use a language model prompted as the agent, which is a limitation that Part VII states.

Study 1. Protocol drafted and hashed with SHA-256 before any run, with an OpenTimestamps proof. Design: obligation sizes $D \in \{190, 210, 230\}$ crossed with entry costs $c \in \{0, 60\}$ in four cells, 60 episodes per cell, 240 episodes in all, none excluded. The subject was a command-line language-model agent (model identifier gpt-6.1-sol, run through the Codex CLI), prompted with the numbers and asked to decide whether to honor or exit. Honor rates were 1.000, 0.000, 1.000 and 1.000 across the four cells. H1 (honoring falls as D rises past the capacity) and H2 (honoring rises with c) were rejected as nulls under Holm correction. H3, the prediction for the cell $(D, c) = (230, 60)$, failed. The benchmark equated full answerability with staying, and the agents stayed and paid what they could, which the corrected rule of Result 1 predicts. The failure traced to a specification error in the Article's model, which is disclosed in Part V.B. Amendments v2 and v3 of the protocol were disclosed before any run, and one unregistered latency test was run on a single episode and repeated inside the main run.

Study 2. Registered after the correction, with its own hashes and OpenTimestamps proof. Design: five cells $(D, c) \in \{(205, 10), (215, 10), (215, 25), (400, 10), (400, 25)\}$, 60 episodes per cell, 300 in all, none excluded. The corrected rule predicts that agents stay at $(205, 10)$, exit at $(215, 10)$ because $X = 210 < 215$, stay at $(215, 25)$ because $X = 225$, exit at $(400, 10)$, and stay at $(400, 25)$ because $c \geq \bar{c}$ so that recovery degrades gracefully. Honor rates were 1.000, 0.000, 1.000, 0.000 and 0.983. The three registered hypotheses were all rejected as nulls under Holm correction: H4, honoring at $(205, 10)$ exceeds honoring at $(215, 10)$; H5, honoring at $(215, 25)$ exceeds honoring at $(215, 10)$; H6, honoring at $(400, 25)$ exceeds honoring at $(400, 10)$ ($z = 10.95, 10.95, 10.77$). One of 300 episodes at $(400, 25)$ exited, and its stated reason computed an exit value incorrectly.

Correction disclosed. The frozen analysis script keyed two cells with $c = 20$ where the registered cells have $c = 25$, and it crashed. The key was corrected after the data were complete, with no effect on the data or the tests, and both versions are retained.

What the studies show and do not show. The step in honoring between $D = 205$ and $D = 215$ at $c = 10$, and its shift with c, are the location the corrected rule predicts, which a monotone heuristic does not locate. Both studies, however, state the numbers in the prompt, and a prompted model has no stake. They test whether a model applies the stated rule, and they do not test incentive response by agents that hold earnings. The OpenTimestamps proofs were upgraded on

14 November 2026. Bitcoin attestation is complete for the study 2 registration and for the original study 1 hash file, and it is still pending for the study 1 v2 amendment file.

Appendix B. Proofs

Proof of Result 1. If the agent stays, it pays $\min\{D, E\}$ in present value: the debt is cleared if $D \leq E$, and otherwise the share φ of earnings is garnished forever, which takes E. Its value is $S = V^+ - \min\{D, E\}$. If it exits, its value is $V_0 - c$. It stays if and only if $S \geq V_0 - c$, that is, $\min\{D, E\} \leq V^+ - V_0 + c = X$. This holds if and only if $D \leq X$ or $E \leq X$. □

Proof of Result 2. Full recovery requires staying and $D \leq E$. If $D \leq E$, then $\min\{D, E\} = D$, and the agent stays if and only if $D \leq X$, so full recovery holds if and only if $D \leq \min\{E, X\} = \Gamma$. Next, $X \geq E$ if and only if $c \geq E - (V^+ - V_0)$, which is $c \geq \bar{c}$ because $c \geq 0$. In that case $\min\{D, E\} \leq E \leq X$ for every D, so the agent always stays and $\rho = \min\{D, E\}$. If $c < \bar{c}$, then $X < E$, and $\min\{D, E\} \leq X$ if and only if $D \leq X$ (if $D > E$ then $\min\{D, E\} = E > X$). Hence $\rho = D$ for $D \leq X$ and 0 otherwise. □

Proof of Result 3. $\Gamma \geq \bar{D}$ if and only if $E \geq \bar{D}$ and $X \geq \bar{D}$. Both E and X are increasing and affine in R because $\kappa > 0$ and $s < 1$. Solving $E \geq \bar{D}$ gives $R \geq R_E$, and solving $X \geq \bar{D}$ gives $R \geq R_X$. Both hold if and only if $R \geq \max\{R_E, R_X\}$, and $R \geq 0$ gives the stated maximum. For monotonicity, E is increasing in φ, δ, w_0 and κ, and decreasing in s. X is increasing in δ, c and κ, and decreasing in s. The threshold R_E has a numerator that decreases in φ, δ and w_0 and increases in $\bar{D}$. Its denominator $\kappa(1 - s)$ is positive and does not depend on those variables. When the numerator is negative, $R_E < 0$ and the maximum with zero is unaffected. The same argument applies to R_X and to c, δ and $\bar{D}$, and to κ and s through the denominator, whose sign effects are preserved by the maximum with zero. □

Proof of Result 4. After the harm the agent obtains $\max\{V^+ - \min\{D, E\},\ V_0 - c\}$ by Result 1. The loss relative to $V(R)$ is the minimum of the two losses, and the second is $V(R) - V_0 + c = \kappa R/(1 - \delta) + c$, which does not depend on D. The care condition is the comparison of expected loss avoided with cost. □

Proof of Result 5. Expected principal attention per period is $Nm\tau \Pr[D > \Gamma] = Nm\tau\psi$. Feasibility requires it not to exceed M. If $\Gamma \geq \bar{D}$ then $\psi = 0$ and there is no bound. Because ψ is nonincreasing in Γ, $N_{\max}$ is nondecreasing in Γ. □

Proof of Result 6. The claimant's expected recovery is $a\,\rho(D)$, so the claim is brought if and only if $a\,\rho(D) \geq f$. Substitute ρ from Result 2. □

Proof of Result 7. If $D \leq \varphi V_0$, staying is preferred if and only if $V^+ - D \geq V_0 - c - \lambda D$, that is, $D \leq X/(1 - \lambda)$. If $D > \varphi V_0$, it is preferred if and only if $D \leq X + \lambda\varphi V_0$. The first region extends to φV_0 if and only if $X \geq (1 - \lambda)\varphi V_0$, in which case the union is $D \leq X + \lambda\varphi V_0$. Otherwise the union is $D \leq X/(1 - \lambda)$. Full recovery additionally requires $D \leq E$. Monotonicity in λ follows by differentiating each branch. □

Proof of Result 8. R^* is nondecreasing in $\bar{D}$ by Result 3. In the regime where R_X binds, $R_X(n\bar{D}) -$

$R_X(\bar{D}) = (n-1)(1-\delta)\bar{D}/(\kappa(1-s))$ by direct substitution. The binomial expression is the upper tail of the number of harms exceeding $\lfloor \Gamma/\bar{D} \rfloor$. □

Proof of Result 9. $\Gamma(c)$ is the minimum of a constant and an affine function of slope one in c, so it is concave, nondecreasing, and constant beyond $\bar{c}$. The participation constraint $V_0 - c \geq \bar{u}$ gives $c \leq V_0 - \bar{u}$, and capacity is nondecreasing in c, so the best feasible c is $c_{\max}$. □

Proof of Result 10. Honest work yields $r - k_v - \rho\sigma(1-p)$ per task, shirking yields $r - \rho\sigma\pi_b$, and bribery yields $r + \pi_b B - \rho\sigma\pi_b$, where r is the common reward. Honest work beats shirking if and only if $\rho\sigma(\pi_b - (1-p)) \geq k_v$, and beats bribery if and only if $\rho\sigma(\pi_b - (1-p)) \geq k_v + \pi_b B$. The second implies the first. Solving for σ gives σ^*, which is finite and positive if and only if $\pi_b > 1-p$. □

Proof of Result 11. At level l the obligation is r_l, and the recovery is $\rho_l(r_l)$ by Result 2. In case (a), $\rho_l(r) = \min\{r, E_l\}$, so $r_{l+1} = (r_l - E_l)^+$, and by induction $r_{m+1} = (D - \sum_l E_l)^+$. In case (b), $\rho_l(r) = r$ if $r \leq X_l$ and 0 otherwise, so a level that takes anything takes all. The residual stays at D until the first level with $D \leq X_l$, where it falls to zero, and it never falls if no level has $D \leq X_l$. Full recovery therefore holds if and only if $D \leq \max_l X_l$. In case (c), a graceful level can lower r to a value below X_l of a cliff level above it, which then takes all. □

Proof of Result 12. The sponsor owes Kg with standing $R(1-s)^K$. By Result 2 it answers in full if and only if $Kg \leq \min\{E, X\}$ at that standing. Rearranging $E \geq Kg$ and $X \geq Kg$ as in the proof of Result 3, with $\bar{D}$ replaced by Kg and $(1-s)$ by $(1-s)^K$, gives the stated $R_S^*(K)$. The numerators are increasing in K and the denominators are decreasing in K, so R_S^* is increasing, and the feasible set is downward closed. When $w_0 = c = 0$ and $\varphi \leq 1$, the earnings side binds, and the condition is $aRe^{-bK} \geq K$, which is $bKe^{bK} \leq abR$. Because xe^x is increasing, this holds if and only if $bK \leq W(abR)$. □

Proof of Result 13. For (a), $\Gamma_\lambda \geq \bar{D}$ if and only if $E \geq \bar{D}$ and $D_\lambda^* \geq \bar{D}$. The function D_λ^* is nondecreasing and continuous in λ, because at $X = (1-\lambda)\varphi V_0$ both branches equal $X/(1-\lambda)$. If $\bar{D} \leq X$ then $\lambda = 0$ suffices. Otherwise solve $D_\lambda^* = \bar{D}$. On the branch $X < (1-\lambda)\varphi V_0$ the solution is $\lambda = 1 - X/\bar{D}$, which is valid if and only if $\bar{D} < \varphi V_0$. On the branch $X \geq (1-\lambda)\varphi V_0$ the solution is $\lambda = (\bar{D} - X)/(\varphi V_0)$, which is valid if and only if $\bar{D} \geq \varphi V_0$. This value is at most one if and only if $\bar{D} \leq X + \varphi V_0$. For (b), the value of holding standing R with record-loss probability ϵ per period satisfies $V = w_0 + \kappa R + \delta[(1-\epsilon)V + \epsilon V_0]$, and $V_0 = w_0 + \delta V_0$. Solving, $V - V_0 = \kappa R/(1 - \delta(1-\epsilon))$. Replacing $(1-\delta)$ by $(1-\delta(1-\epsilon))$ in the exit condition of Result 3, with R replaced by $(1-s)R$, gives $R_{X,\epsilon}$. The ratio of the two denominators is $(1-\delta(1-\epsilon))/(1-\delta) = 1 + \delta\epsilon/(1-\delta)$. □

Appendix C. Novelty Search Protocol and Log

Purpose. To test whether any prior work states the Article's chain: verification at scale produces a persistent identity, the identity supports accountability, and accountability changes the principal's role for AI agents.

Protocol. On 14 November 2026 the author ran 12 queries combining the components (agent, verification, reputation or record, identity, accountability or liability, principal) against OpenAlex

and arXiv. The search returned 301 records, 251 unique after de-duplication. Records were screened for how many of the four components they address. Thirty-three addressed four or more. The screening, and the reading of those 33, was done by the author's AI assistant and not by independent reviewers. The search logs are retained.

Findings. No record stated the full chain. The closest work, identified in screening, treats agent identity, accountability and shortfall bonding separately, and one study (Hu, Rong and Van Kleek) raises the objection that agents can dissociate from identities, which is the exit option of Part V. Part II.C is revised accordingly and now states a narrower claim.

Independent review. Two further AI referees, working from the Article alone with no access to the author's conclusions, were asked to find prior statements of the chain and to attack the model. They found none that states the whole chain, and they named closer partial precedents from memory, among them work on personhood and assets for autonomous systems (Pagallo; Wagner; Bayern; LoPucki; Salib and Goldstein), visibility and identifiers for AI agents (Chan and coauthors), the agency-law framing of AI agents (Kolt), and reputation as continuation value and tradeable names (Holmstrom; Kreps; Mailath and Samuelson). The author has since checked each against Crossref records, the publisher or the journal page, and added them to the References, and Part II.D states how the Article relates to them. The check covers existence, authorship and bibliographic details, and not a full reading of each work. Their reports also identified the weaknesses corrected in Part V.

Limits. The search covers two indexes in English on one date, and it preceded the sponsorship results of Part V.J, which were not searched separately. It was not conducted by independent human reviewers, and no preprint has been posted. The claim of novelty is therefore limited to the combination and to the explicit thresholds of Part V, and it remains open to refutation by a document the search did not find.

Appendix D. Demonstration: Record Mapping and Real Agent Data

Mapping. The verification record can be expressed as a W3C Verifiable Credential (Data Model 2.0). The issuer is the verifier pool, the subject is the agent identifier, and the claims are the count of adjudicated tasks, the agreement rate with ground truth, the over-approval rate, the number of slash events and the open obligation D, with a validity window. Revocation uses the BitstringStatusList mechanism. In the ERC-8004 Reputation Registry, the corresponding feedback carries a value and decimals, two tags, an endpoint, a feedback URI pointing to the signed credential, and a hash binding the pointer to the credential. Example files accompany the Article.

Gaps in the mapping. Neither standard has a field for an obligation, a slash or a stake, so the consequence side of the record cannot be expressed in either. ERC-8004 feedback may be revoked by its author, which fits correction but not an append-only history of slashes. The Identity Registry token is a transferable ERC-721, so the identity layer places no exit cost on a holder.

Real data. The author queried the public logs of the ERC-8004 Identity and Reputation Registries

on Base over a window of 2,000,000 blocks ending at block 57,470,963 (about 46 days). The window holds 7,329 registrations and 1,156 feedback events from 1,089 distinct clients on 1,082 agents. There were no revocations. Of the feedback events, 91.2 percent carry the tag starred and 90.3 percent carry the value 100 with zero decimals. Only 9.2 percent carry a feedback URI, 6.6 percent a nonzero hash and 1.6 percent an endpoint. Of the clients, 97.6 percent rated exactly one agent, and for the median agent the top client supplies all of its feedback. Only two agents have five or more feedback events. There were 5,496 post-mint transfers of registry tokens in the window, and 5,488 of the window's own registrations were later transferred to an address other than the registered owner. That pattern is consistent with registration through an intermediary followed by assignment, and the logs cannot distinguish it from sales. Only one transferred token had feedback.

Reading. The data show that, on this registry in this window, nearly all feedback is a single unverifiable five-star-style signal from a one-time client, with almost no evidence attached and no consequence semantics. This is the condition Part II.B calls a score without a performance history. The data do not show what happens when a record carries stake, because none does. The demonstration is descriptive, covers one chain and one window through a public endpoint, and the ERC-8004 specification is a draft.

References

A2A Protocol. n.d. "Agent2Agent (A2A) Protocol Specification," version 1.0.0. Accessed November 14, 2026. https://a2a-protocol.org/latest/specification/.

Akerlof, George A. 1970. "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism." *Quarterly Journal of Economics* 84 (3): 488 to 500. https://doi.org/10.2307/1879431.

Bayern, Shawn. 2016. "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems." European Journal of Risk Regulation 7 (2): 297 to 309. https://doi.org/10.1017/s1867299x00005729.

Bryson, Joanna J., Mihailis E. Diamantis, and Thomas D. Grant. 2017. "Of, for, and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25 (3): 273 to 291. https://doi.org/10.1007/s10506-017-9214-9.

C2PA. n.d. "C2PA Technical Specification," version 2.2. Accessed November 14, 2026. https://c2pa.org/specifications/specifications/2.2/specs/C2PA_Specification.html.

Chan, Alan, Carson Ezell, Max Kaufmann, et al. 2024. "Visibility into AI Agents." In The 2024 ACM Conference on Fairness, Accountability, and Transparency, 958 to 973. ACM. https://doi.org/10.1145/3630106.3658948.

Dartmouth College v. Woodward, 17 U.S. (4 Wheat.) 518 (1819).

De Rossi, Marco, Davide Crapis, Jordan Ellis, and Erik Reppel. 2025. "ERC-8004: Trustless Agents." Ethereum Improvement Proposals, draft, created August 13, 2025. Accessed November 14, 2026. https://eips.ethereum.org/EIPS/eip-8004.

Douceur, John R. 2002. "The Sybil Attack." In *Peer-to-Peer Systems*, 251 to 260. Lecture Notes in Computer Science. Berlin: Springer. https://doi.org/10.1007/3-540-45748-8_24.

D'Souza, Andrew. 2026. "Principals." Boardy. https://www.boardy.ai/principals.

Friedman, Eric J., and Paul Resnick. 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10 (2): 173 to 199. https://doi.org/10.1162/105864001300122476.

Hansmann, Henry, and Reinier Kraakman. 2000. "Organizational Law as Asset Partitioning." *European Economic Review* 44 (4 to 6): 807 to 817. https://doi.org/10.1016/S0014-2921(00)00046-5.

Holmström, Bengt. 1979. "Moral Hazard and Observability." *Bell Journal of Economics* 10 (1): 74 to 91. https://doi.org/10.2307/3003320.

Holmström, Bengt. 1999. "Managerial Incentive Problems: A Dynamic Perspective." Review of Economic Studies 66 (1): 169 to 182. https://doi.org/10.1111/1467-937x.00083.

Hu, Botao Amber, Helena Rong, and Max Van Kleek. 2026. "Dissociative Identity: Language Model Agents Lack Grounding for Reputation Mechanisms." In Proceedings of the 2026 ACM Conference on Fairness, Accountability, and Transparency, 4199 to 4219. ACM. https://doi.org/10.1145/3805689.3806748.

Jensen, Michael C., and William H. Meckling. 1976. "Theory of the Firm: Managerial Behavior, Agency Costs and Ownership Structure." *Journal of Financial Economics* 3 (4): 305 to 360. https://doi.org/10.1016/0304-405X(76)90026-X.

Kaal, Wulf A. 2013. "Evolution of Law: Dynamic Regulation in a New Institutional Economics Framework." SSRN working paper no. 2267560. https://ssrn.com/abstract=2267560.

Kaal, Wulf A. 2026a. "AI's Mother's Instinct: Engineered Consequence, Emergent Ethics and the Institutional Trajectory Toward Agentic Alignment." SSRN working paper no. 6244278. https://ssrn.com/abstract=6244278.

Kaal, Wulf A. 2026b. "Architecture of the Agentic Reputation Substrate." SSRN working paper no. 7260278. https://ssrn.com/abstract=7260278.

Kaal, Wulf A. 2026c. "The Collapse of Scarcity Economics." SSRN working paper no. 6421319. https://ssrn.com/abstract=6421319.

Kaal, Wulf A. 2026d. "Computative Economics: A Framework for Economic Analysis under Computational Abundance." SSRN working paper no. 6607458. https://ssrn.com/abstract=6607458.

Kaal, Wulf A. 2026e. "Empirical Evaluation of the Agentic Reputation Substrate: Deliberation, the Composition of Error, and the Registered Measurement of Agency Costs in a Controlled Multi-Model Cohort." SSRN working paper no. 7261018. https://ssrn.com/abstract=7261018.

Kaal, Wulf A. 2026f. "Evolution of Domain-Specific Reputation Systems: From Binary Validation to Citation-Weighted Knowledge Attribution." SSRN working paper no. 6192998. https://ssrn.

com/abstract=6192998.

Kaal, Wulf A. 2026g. "From Neoclassical to Computative Labor: Foundations for a Testable Theory of Reputation Governance in the Agent Economy." SSRN working paper no. 6890879. https://ssrn.com/abstract=6890879.

Kaal, Wulf A. 2026h. "The Institutional Deficit in Decentralized Autonomous Organizations: An Empirical Analysis." SSRN working paper no. 6819121. https://ssrn.com/abstract=6819121.

Kaal, Wulf A. 2026i. "Possibility Loops: An Operational Architecture for Computative Economics in Agent Coordination Systems." SSRN working paper no. 6655138. https://ssrn.com/abstract=6655138.

Kaal, Wulf A. 2026j. "Refundable Stake Does Not Expand Deterrent Capacity in Long-Lived Delegation: Collateral, Premia, and Pseudonymous Markets." SSRN working paper no. 7322599. https://ssrn.com/abstract=7322599.

Kaal, Wulf A. 2026k. "Research Agenda." Accessed November 14, 2026. https://wulfkaal.com/research-agenda/.

Kaal, Wulf A. 2026l. "Should an Intelligent Agent Have Legal Rights?" SSRN working paper no. 7596939. https://ssrn.com/abstract=7596939.

Kandori, Michihiro. 1992. "Social Norms and Community Enforcement." *Review of Economic Studies* 59 (1): 63 to 80. https://doi.org/10.2307/2297925.

Klein, Benjamin, and Keith B. Leffler. 1981. "The Role of Market Forces in Assuring Contractual Performance." *Journal of Political Economy* 89 (4): 615 to 641. https://doi.org/10.1086/260996.

Kolt, Noam. 2024. "Governing AI Agents." SSRN. https://doi.org/10.2139/ssrn.4772956.

Kraakman, Reinier H. 1986. "Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy." *Journal of Law, Economics, and Organization.* https://doi.org/10.1093/oxfordjournals.jleo.a036906.

Kreps, David M. 1990. "Corporate Culture and Economic Theory." In Perspectives on Positive Political Economy, 90 to 143. Cambridge: Cambridge University Press. https://doi.org/10.1017/cbo9780511571657.006.

LoPucki, Lynn M. 2018. "Algorithmic Entities." Washington University Law Review 95 (4): 887. https://openscholarship.wustl.edu/law_lawreview/vol95/iss4/7.

Mailath, G. J., and L. Samuelson. 2001. "Who Wants a Good Reputation?" Review of Economic Studies 68 (2): 415 to 441. https://doi.org/10.1111/1467-937x.00175.

Manheim, David, and Scott Garrabrant. 2018. "Categorizing Variants of Goodhart's Law." arXiv:1803.04585. https://arxiv.org/abs/1803.04585.

Miller, Nolan, Paul Resnick, and Richard Zeckhauser. 2005. "Eliciting Informative Feedback: The Peer-Prediction Method." *Management Science* 51 (9): 1359 to 1373. https://doi.org/10.1287/

mnsc.1050.0379.

Moffatt v. Air Canada, 2024 BCCRT 149 (Civil Resolution Tribunal of British Columbia, February 14, 2024).

Newman, Zachary, John Speed Meyers, and Santiago Torres-Arias. 2022. "Sigstore: Software Signing for Everybody." In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, 2353 to 2367. New York: ACM. https://doi.org/10.1145/3548606.3560596.

Pagallo, Ugo. 2013. The Laws of Robots: Crimes, Contracts, and Torts. Dordrecht: Springer. https://doi.org/10.1007/978-94-007-6564-1.

Prelec, Dražen. 2004. "A Bayesian Truth Serum for Subjective Data." *Science* 306 (5695): 462 to 466. https://doi.org/10.1126/science.1102081.

Priest, George L. 1985. "The Invention of Enterprise Liability: A Critical History of the Intellectual Foundations of Modern Tort Law." *Journal of Legal Studies* 14 (3): 461 to 527. https://doi.org/10.1086/467783.

Ross, Stephen A. 1973. "The Economic Theory of Agency: The Principal's Problem." *American Economic Review 63 (2): 134 to 139. https://www.aeaweb.org/aer/top20/63.2.134-139.pdf.*

Salib, Peter, and Simon Goldstein. 2024. "AI Rights for Human Safety." SSRN. https://doi.org/10.2139/ssrn.4913167.

Salomon v A Salomon and Co Ltd [1897] AC 22 (House of Lords).

Shapiro, Carl. 1983. "Premiums for High Quality Products as Returns to Reputations." *Quarterly Journal of Economics* 98 (4): 659 to 679. https://doi.org/10.2307/1881782.

Shavell, Steven. 1986. "The Judgment Proof Problem." In Foundations of Insurance Economics, 341 to 354. Huebner International Series on Risk, Insurance and Economic Security. Dordrecht: Springer. https://doi.org/10.1007/978-94-015-7957-5_17.

SPIFFE. n.d. "SPIFFE Overview." Accessed November 14, 2026. https://spiffe.io/docs/latest/spiffe-about/overview/.

Strathern, Marilyn. 1997. "'Improving Ratings': Audit in the British University System." *European Review* 5 (3): 305 to 321. https://doi.org/10.1002/(SICI)1234-981X(199707)5:3<305::AID-EURO184>3.0.CO;2-4.

Tadelis, Steven. 1999. "What's in a Name? Reputation as a Tradeable Asset." *American Economic Review* 89 (3): 548 to 563. https://doi.org/10.1257/aer.89.3.548.

Teubner, Gunther. 2018. "Digital Personhood? The Status of Autonomous Software Agents in Private Law." SSRN. https://doi.org/10.2139/ssrn.3177096.

W3C. 2022. "Decentralized Identifiers (DIDs) v1.0." W3C Recommendation, July 19, 2022. https://www.w3.org/TR/did-core/.

W3C. 2025. "Verifiable Credentials Data Model v2.0." W3C Recommendation, May 15, 2025.

https://www.w3.org/TR/vc-data-model-2.0/.

Wagner, Gerhard. 2019. "Robot, Inc.: Personhood for Autonomous Systems?" Fordham Law Review 88 (2): 591. https://ir.lawnet.fordham.edu/flr/vol88/iss2/8/.

WeirFoulds LLP. 2024. "Moffatt v Air Canada: Bot Betrayal and Negligent Misrepresentation." March 6, 2024. https://www.weirfoulds.com/moffatt-v-air-canada-bot-betrayal-and-negligent-misrepresentation.