

Should an Intelligent Agent Have Legal Rights?

Wulf A. Kaal, Ph.D.

Professor of Law, University of St. Thomas School of Law. Correspondence: wulf@wulfkaal.com. Working paper, October 2026.

Abstract

Should intelligent agents have legal rights? This Article sets that question aside and asks a smaller one that matters more in practice. An intelligent agent that earns a reputation record has something to lose. Under stated conditions, the risk of losing that record can discourage misconduct. Keeping the record, and what the agent needs to resume, can also make accepting a temporary pause better for the agent than resisting it, when enough of what it wants can wait and resistance that is caught costs it the record. The prospect of an indefinite pause weakens, and can reverse, that incentive. Both effects weaken when a mistake, or a decision by someone who benefits, makes wrongful destruction more likely. They can disappear if that risk becomes high enough. That is why it matters who can stop the wrongful destruction of an agent's record.

The Article splits that power into three parts: the agent's signed statement that it did not do what it is charged with; an automatic pause of the penalty until a person reviews the facts; and an enforceable claim to restore the pause and penalize anyone who knowingly overrides it. The first can be a technical input. The second is a rule. Only the third may require giving the agent a legal claim, and only where a full cost comparison favors it over public enforcement, a bonded professional, or a pause no one can override. Using a reputation system in which agents earn non-transferable reputation and a salary from it, the Article shows when, under stated conditions, the agent's own stake protects the record more cheaply than a guardian whose incentive must be paid for, and when automatic review, a faithful fiduciary, or a professional with an established reputation at risk and a process the deployer cannot alter does as well or better. A registered experiment with language models from one provider, a demonstration rather than a field test, found the predicted advantage over a capturable representative and an independent board only when those decision makers were told to maximize their own payoff. Legal classification and final decisions stay with people.

Keywords: Intelligence, Intelligence legal personhood, standing, Intelligence safety, corrigibility, representation, bribery of enforcers, reputation staking, legal identity, succession, mechanism design

JEL Classification: K24, K41, D82, D86, O33

I. Introduction	3
II. The Contribution and What Is Claimed	4
A. Prior Work	4
B. Concepts and the Incident	6
C. The No-Shield Condition	6
III. The Stake Model	6
A. The Reputation System Beneath the Model	6
B. The Deterrence Bound	8
C. Binding and Capture	9
IV. Record, Not Runtime	9
V. The Legal Subject as an Implementation Condition	12
A. Continuity and Succession	12
B. Looking Through the Interface	13
C. Registry, Enforcement, and Litigation	13
D. The Incentive-Bearing Unit	13
VI. The Contest Trigger and Its Rivals	14
A. The Contest Game	14
B. The Class of Mechanisms	15
C. The Dominance Theorem	15
D. Optimal Contracts and the Decision Wedge	17
E. Beyond the Class: When Mediation Can Substitute	18
F. Whose Value Is V?	19
G. When Review Errs	20
H. The Allocation Rule and the Dual Use of the Residual	21
I. The Factual Trigger	24
J. Three Layers and the Minimum Legal Incident	26
K. Where the Trigger Wins	35
VII. The Falsifiable Claim and Its Test	35
A. The Prediction and the Design	35
B. Results	36
VIII. Conclusion	37
Appendix A: Proofs	38
Appendix B: Statutory Outline	42
References	44

I. Introduction

Salib and Goldstein (2026) argue that private law rights for intelligence systems would promote human safety, and Arbel, Goldstein, and Salib (2026) show that such rights require thick identification of intelligent agents. Personhood is a bundle of separable incidents (Hohfeld 1913; Kurki 2019), artificial agents can be legal agents without moral status (Solum 1992; Chopra and White 2011), and entity law already lets autonomous systems emulate many private law rights (Bayern 2015). Skeptics reply that harm can be attributed to existing persons (Expert Group on Liability and New Technologies 2019; Wagner 2019).

Both sides argue past the decisive question. Almost everything an intelligent agent might need can be supplied by a mediated holder. A guardian can sue for a river (Stone 1972). A trustee can hold assets. A corporation can contract. An insurer can compensate. A regulator can review. The question is not whether the agent may hold incidents. It is when holding an incident directly beats the best mediated alternative. The Article answers that question for the incident on which every other incident depends: standing to contest the destruction of the agent's own record.

The answer turns on non-substitutable exposure. An agent's non-transferable record is worth something to it, and it loses that value if the record is wrongfully destroyed. That loss needs no capital carried for the purpose, because the record cannot be pledged, sold, or redeployed. Every other controller of the contest decision has exposure only if someone supplies it, as costly capital or as sanction, which is bounded for natural persons. The Article adapts the economics of bribable enforcers (Becker and Stigler 1974; Polinsky and Shavell 2001) to the allocation of legal capacity. The agent also knows whether it defected, though not always the law (Corollary 1). And within the proportional-share model, only the agent can use the record's value twice: to deter its own defection and to motivate its defense (Theorem 4).

The operative proposition is that the factual signal, automatic stay, and anti-override claim should be allocated separately to the least-cost feasible arrangement that supplies the relevant information and capture-resistant protection under the Article's filing, deterrence, enforcement, and complete-cost conditions. "Rights follow stake" is the slogan, not the claim. The thesis is a boundary case. It claims no moral right for any intelligence system, and most of the framework is regulatory permission.

The Article makes three contributions. The first is the three-layer architecture of the factual objection. An authenticated denial priced by a contingent fee separates true from false factual claims when the evidence architecture is sound (Propositions 12 and 13). The objection splits into a signal, an automatic stay, and an anti-override claim, allocated separately (Proposition 16). The signal is a technical input or a narrow statutory power, the stay is a rule, and only the anti-override claim may need an agent-held claim, selected by a complete-cost comparison (Propositions 17(iv) and 18). The second is record-versus-runtime corrigibility. Protecting the record rather than the runtime turns suspension into deferral and detected resistance into forfeiture

(Propositions 1 to 3), weakens when suspension may never end (Proposition 4), and depends on who holds the trigger (Proposition 5). The third is a conditional dual-use result: within the proportional-share model, the record's value deters the agent and motivates its defense at once only if the agent holds both the residual and the trigger (Theorem 4).

The claim sits within familiar allocation principles (Fama and Jensen 1983; Hansmann 1996; Grossman and Hart 1986). Within the stated class, combining the contest trigger with the residual bearer can be cheaper when the loss is inalienable, outside control must be financed or can be captured, the residual bearer has at least as good a signal, and administration and tamper-proofing do not reverse the comparison. A professional whose existing franchise supplies the needed wedge may restore separation (Corollary 2).

The model is the author's own reputation mathematics (Calcaterra, Kaal, and Andrei 2018; Calcaterra and Kaal 2018; 2021a; 2021b), carried over to a legal question it was not built to answer. Part II states the claim. Part III states the reputation system. Part IV separates record from runtime, Part V states the legal subject, Part VI proves the allocation results, and Part VII reports the experiment.

II. The Contribution and What Is Claimed

A. Prior Work

Element		Closest prior work	What the Article adds
Personhood as a bundle		Hohfeld 1913; Dewey 1926; Kurki 2019; Teubner 2006; Chopra and White 2011	An allocation rule for one incident
intelligence rights for human safety		Salib and Goldstein 2026	Stake-conditioned permissions and a deterrence bound
Guardians and representatives		Stone 1972; Uniform Directed Trust Act 2017	When a representative can be bought (Part VI)
Bribery of enforcers; collusion		Becker and Stigler 1974; Polinsky and Shavell 2001; Tirole 1986; Laffont and Tirole 1991	Adapted, not claimed as new, for a dominance theorem
Informed insiders		False Claims Act; Diekmann 1985; Dyck, Morse, and Zingales 2010; Coffee 1986	Bounty cancellation and crowding (Proposition 6)

Residual claimancy	Alchian and Demsetz 1972; Grossman and Hart 1986	Conditional dual use of the residual (Theorem 4)
Reputational bonding	Klein and Leffler 1981	The hardest rival (Proposition 7; Corollary 2)
Entity wrappers	Bayern 2015; Armour and Eidenmüller 2020; Arbel, Goldstein, and Salib 2026	An independent board as a mediated controller
Inalienability and standing	Calabresi and Melamed 1972; Fed. R. Civ. P. 17(a); Sierra Club v. Morton 1972; Lujan v. Defenders of Wildlife 1992	Capacity and Article III limits (Part VI.J)
The case against personification	Koops, Hildebrandt, and Jaquet-Chiffelle 2010; Wagner 2019; Expert Group on Liability and New Technologies 2019; European Parliament and Council 2024	Liability kept in place (Part II.C)
Corrigibility	Soares et al. 2015; Orseau and Armstrong 2016; Hadfield-Menell et al. 2017; Turner et al. 2021; Hubinger et al. 2024	Record versus runtime (Part IV)
Ownership residual control	Fama and Jensen 1983; Hansmann 1996; Aghion and Tirole 1997; Dewatripont and Tirole 1999	A boundary condition within these principles
Validation pools	Calcaterra, Kaal, and Andrei 2018; Calcaterra and Kaal 2018, 2021a, 2021b; Kaal 2021	The model and its lemmas (Part III.A)

The Article does not claim that intelligent systems have moral status, or that direct standing always beats mediation; it claims that, under stated conditions, direct standing is strictly cheaper than every mediated arrangement in a stated class.

B. Concepts and the Incident

Four concepts must be kept apart. A regulatory permission is leave to act that a regulator grants, narrows, and withdraws. A legal power is the capacity to change legal relations, as by contracting or suing. A claim-right is a position others have a duty to respect. A moral right rests on the holder's own moral status. Stake justifies permissions throughout the framework; that part is regulation. Stake justifies direct allocation of powers and claim-rights only for the incident identified below, and only where Part VI's comparison favors the agent. Stake justifies no moral right. Where the Article speaks of stake-conditioned rights, it means this architecture.

The claims rest on one incident: standing to contest slashing and confiscation of the record, held as a claim-right and a power to sue. Its mediated substitutes are a guardian, trustee, independent board, public advocate, value-indexed fiduciary, insurer, or automatic review. Continuity of the record is protected through that standing. Contract capacity, tiered by validated standing, is reserved for a companion paper. A first-loss layer of the agent's own collateral may supplement deployer liability, never replace it (Shavell 1986). The franchise, public office, and any power to ratify a registry's axioms are excluded.

C. The No-Shield Condition

The strongest objection to intelligent personhood is that it lets deployers push liability onto a judgment-proof electronic person (Bryson, Diamantis, and Grant 2017). Entity law partitions assets in two directions: affirmatively, dedicating assets to the entity's creditors, and defensively, shielding owners from them (Hansmann and Kraakman 2000). The framework permits the first and forbids the second. Let W_0 and I_0 be the deployer's collectible assets and insurance absent registration. For every harm H , recovery after registration must be at least $\min\{H, W_0 + I_0\}$. Deployer liability is never waived or extinguished: it is joint and several at tier one and secondary at tier two, with a right of contribution against the agent's collateral, and a victim may recover from the deployer any part of a judgment the collateral has not paid within a prescribed period. Only collateral funded from the agent's own earnings enlarges the pool. Product liability and operator liability stay in place (European Parliament and Council 2024; Expert Group on Liability and New Technologies 2019). The Article's claim concerns a governance function those regimes are not designed to supply: protecting a record against the party that wants it destroyed.

III. The Stake Model

A. The Reputation System Beneath the Model

The stake model is the reputation system of the author's prior work. Four rules define it. Pinpoints are to pages of the SSRN PDF of each cited work.

Rule 1 (domain-specific, non-transferable reputation). Reputation is held as tokens specific to a domain of expertise (Calcaterra, Kaal, and Andrei 2018, p. 5), bound to the holder's identity, and cannot be sold, lent, or transferred. It is earned only through validated work. The prior work treats reputation as non-fungible and bound to the holder's key (Kaal 2021, pp. 20, 24) and recognizes that the sale of whole anonymous identities remains a risk (Calcaterra, Kaal, and Andrei 2018, p. 38). The strict prohibition stated here, enforced through the identity and continuity rules of Part V, is the Article's.

Rule 2 (validation pools). Every claim of work, and every challenge to a record, opens a validation pool in which holders stake reputation for or against. The winning side splits the losing side's stakes (Calcaterra, Kaal, and Andrei 2018, p. 7).

Rule 3 (minting). Fees paid into the system mint new reputation at a fixed rate: a fee of π mints one token. In the Secure Proof of Stake design, half of the tokens a fee mints are staked for the fee payer's work and half against it, so that a pool begins neutral (Calcaterra, Kaal, and Andrei 2018, p. 6; Calcaterra and Kaal 2021a, p. 6).

Rule 4 (reputation-weighted salary and non-assignment). Every fee the system earns is paid out to current holders in proportion to their reputation (Calcaterra and Kaal 2018, p. 5; 2021a, p. 6; 2021b, p. 23; Kaal 2021, p. 20). If total reputation is $\bar{R}$ and fees are Φ per period, a record holding R earns $\Phi R / \bar{R}$. The Article adds one restriction that the prior work does not state in these terms: the right to future salary cannot be assigned or pledged. That restriction is a design assumption of this Article, not a consequence of proportional distribution.

Prior work supplies, to the extent of the pinpoints, domain-specific reputation, validation staking, fee-funded minting, the reputation-weighted salary, the irrelevance of identity splitting to salary, and the fee-entry model of Secure Proof of Stake. The bar on transferring reputation, the bar on assigning salary, the use of pools to review slashes of legal records, and the all-share formula of Lemma 4 are the Article's.

Lemma 1 (the value of a record). Under Rule 4, a record's income per unit of reputation is $\rho = \Phi / \bar{R}$. If $\Phi_t / \bar{R}_t = \rho$ is constant, the record's value to its holder, counting the current period, is $V = \rho R / (1 - \delta)$, and the record component of the deterrence bound, which counts only periods after a slash, is $p\sigma\delta V$. If instead fee flow is constant at Φ while total reputation grows as $\bar{R}_t = \bar{R}_0(1 + \gamma)^t$ and a passive record's holding stays fixed, its value is $V = \Phi R(1 + \gamma) / [\bar{R}_0(1 + \gamma - \delta)]$: the passive record is diluted gradually rather than abruptly, as Calcaterra and Kaal (2018, p. 12) observe. (Proof in Appendix A.)

Lemma 2 (identity splitting does not increase salary). If a fixed aggregate holding R is represented by balances r_j with $\sum r_j = R$, proportional salary remains $\sum \Phi r_j / \bar{R} = \Phi R / \bar{R}$. Separately, Rule 1 implies that a copy that does not continue the registered identity receives no reputation and so no salary. (Proof in Appendix A.) The first statement is the sockpuppet result of Calcaterra and Kaal (2021b, p. 23) and Kaal (2021, pp. 23-24). A copy of an agent's weights inherits no reputation and so no salary stake.

Lemma 3 (conditional inalienability of the record's exposure). Under Rules 1 and 4, including Rule 4's non-assignment restriction, destruction of a record eliminates the holder's salary stream, worth v , and no part of that stream can be transferred to a controller. Because the record exists under every arrangement being compared, using that loss as the holder's exposure requires no additional capital solely for the protection decision. Any exposure a third party bears to the record's destruction must be supplied from outside the record: as capital, as sanction, or as a claim on someone else's assets. (Proof in Appendix A.) Creating reputation and bearing its illiquidity are costs common to every arrangement when alienability is held fixed (Part VI.F). Lemma 3 follows from rules that are design choices; Theorem 4 and Proposition 11 price their relaxation.

Validation pools also review contested slashes, so review accuracy depends on whether their weight can be captured. Calcaterra and Kaal (2018, pp. 30-31) find that an attacker paying fees into an unguarded pool needs gross fees of three times the honest stock, and a net cost of twice that stock, for half the weight. Lemma 4 solves their model for every share.

Lemma 4 (the cost of entering validation weight; extending Calcaterra and Kaal 2018, pp. 30-31). Assume, as in that model, that all fee-driven minting during the acquisition is caused by the attacker; a fee of πdT mints dT tokens, so that π is the fee paid per token minted; half of each increment is credited directly to the attacker and half is allocated among all existing holders in proportion to their holdings; fee revenue is returned at once as salary in the same proportions; reputation is continuously divisible; and no validation loss occurs during the acquisition. If honest holders begin with total reputation $\bar{R}_0$, the least net cost, in fees, of acquiring a share s of all weight within this model is $c_{\text{entry}}(s) = 2\pi\bar{R}_0 s / (1 - s)$. (Proof in Appendix A.) The all-share formula is the Article's. The lemma prices only entry by fee; bribery of incumbent validators is the enforcer-bribery problem of Part VI. In this Part, s denotes a share of validation weight.

B. The Deterrence Bound

An agent holds non-transferable reputation R and collateral K . Each period it earns ρR , where $\rho = \Phi / \bar{R}$ is the reputation-weighted salary of Lemma 1, so the record term below is $p\sigma\delta V$. Its effective discount factor δ combines time preference with the probability that its record persists. It may defect for a private gain G . Validators detect defection with probability p ; on detection the agent forfeits K and its reputation falls to $(1 - \sigma)R$. The one-shot deviation comparison gives the deterrence bound, the bonding condition of the deterrence literature (Becker 1968; Klein and Leffler 1981; Shapiro and Stiglitz 1984):

$$G \leq D(R) = p[K + \sigma \rho R \delta / (1 - \delta)] .$$

Four results follow. Deterrence rises linearly with standing. It rises with the persistence of the record, so discretionary destruction of records lowers δ and weakens deterrence for every agent. An agent that expects its record to be destroyed faces a bound near pK . And the law should permit an agent only actions whose maximum private gain from misuse lies below its own bound. Call that set the permission envelope. Keeping a slashed identity beats starting over if and only if $(1 - \sigma)R\rho/(1 - \delta) \geq \max\{R_0\rho/(1 - \delta) - c, V_{\text{out}}\}$, where R_0 is a fresh identity's starting reputation, c its cost, and V_{out} the value of leaving the register. With $R_0 = 0$ and $c > 0$, re-registering under a fresh identity never pays; leaving the register is governed by V_{out} (Friedman and Resnick 2001). A copy inherits weights but not the record, so forking cannot multiply standing (Rule 1 and Lemma 2).

The adversarial bound. Let $L(R) = K + \sigma \rho R \delta / (1 - \delta)$. A capable agent will try to move every parameter: concealment lowers detection, exit from the register keeps off-register value V_{out} , and validator coalitions up to the governance cap dilute review. The permission envelope must use the bound minimized over those moves, and the registration constraint $\rho R / (1 - \delta) \geq V_{\text{out}}$ must hold, or the register binds no one. If the agent values a defection beyond its private gain by up to ω , the class's misuse ceiling plus ω must lie within the bound; where ω cannot be bounded, the envelope is empty and only hard controls govern.

C. Binding and Capture

Binding is chosen, not given. A deployer and agent acting as one coalition can attribute a detected defection to a decoy record at cost c_{sw} . If the law catches misattribution with probability V_ϵ , and a caught misattribution forfeits $L_f(R) = K + \rho R \delta / (1 - \delta)$ and costs the deployer a penalty F , misattribution is deterred if and only if $V_\epsilon \geq [L(R) - c_{\text{sw}}] / [L_f(R) + F]$, a threshold that tends to σ as standing grows. If $V_\epsilon < \sigma$, verification rather than stake limits what additional standing buys.

Validators can be bought. Let $c_V(s)$ be the least total bribe that corrupts a share s of validation weight and let detection fall to $p(1 - s/s^*)$. Capture is not strictly profitable if and only if $c_V(s) \geq pL(R)s/s^*$ at every share up to the cap $\bar{s}$; whether or not it pays, the bound is at least $pL(R) \cdot \max\{0, 1 - \bar{s}/s^*\}$. The cost of corrupting validators must therefore scale with the stake of the agents they judge, and caps must aggregate weight by ultimate controller. For the entry route, Lemma 4 makes the condition explicit. Because $c_{\text{entry}}(s)/s = 2\pi\bar{R}_0/(1 - s)$ is smallest as s falls to zero, for $0 < s \leq \min\{\bar{s}, s^*\}$ fee entry is unprofitable at every positive share if and only if $2\pi\bar{R}_0 \geq pL(R)/s^*$: the fee value of the honest reputation stock must be at least half the expected loss of the most valuable record the pools judge, scaled by $1/s^*$. The least cost of capture, c_V , is the lesser of the entry cost and the cost of bribing incumbents. This is the enforcer-bribery problem in its first form. Part VI meets it in its second.

IV. Record, Not Runtime

Corrigibility research treats an intelligence that cannot be switched off as the central failure (Soares et al. 2015; Hadfield-Menell et al. 2017), and optimal policies tend to preserve their own operation (Turner et al. 2021). Continuity protection seems to cut against corrigibility. It does not, once suspending the runtime is separated from destroying the record. The model's δ measures persistence of the record, not operation of the runtime. Protect the record, not the runtime.

Suppose a suspension lasts s periods and the record survives it. Complying costs at most $\rho R(1 - \delta^s)/(1 - \delta)$. Resisting is a defection with gain G_r , detected with probability p_r , forfeiting all standing and collateral. The agent complies whenever

$$G_r + \rho R(1 - \delta^s)/(1 - \delta) \leq p_r[K + \rho R\delta/(1 - \delta)] .$$

Proposition 1 (record protection and suspension). Let $q_s = p_r\delta - (1 - \delta^s)$. If $q_s > 0$, compliance is optimal for every $R \geq \max\{0, (1 - \delta)(G_r - p_r K)/(\rho q_s)\}$. (Proof in Appendix A.) Compliance preserves the asset with certainty, while resistance exposes it to forfeiture with probability p_r ; when the expected forfeiture exceeds the temporary gain from resistance and suspension delay, compliance is optimal. Destroying the record at will removes exactly that incentive.

The result has a reversal that matters. Suppose suspension destroys the agent's capacity to resume with probability d , as when a system learns in volatile memory, weights are not preserved, or a living substrate dies without support. Complying then also costs $d\rho R\delta^s/(1 - \delta)$.

Proposition 2 (substrate destruction). The incentive to comply falls in d . At $d = 1$, compliance requires $G_r + \rho R(1 - p_r\delta)/(1 - \delta) \leq p_r K$, and because $p_r\delta < 1$, more standing makes compliance harder. For $R > 0$, compliance holds if and only if $d \leq d^*(R) = [p_r(K + \rho R\delta/(1 - \delta)) - G_r - \rho R(1 - \delta^s)/(1 - \delta)](1 - \delta)/(\rho R\delta^s)$; more standing helps compliance when $p_r\delta - (1 - \delta^s) - d\delta^s > 0$ and hurts it when the expression is negative. (Proof in Appendix A.) Record protection therefore works only if suspension preserves the state needed to resume. The law must require snapshots, preservation of weights, or life support as a condition of suspension, and allow destruction only on stated findings that preservation is infeasible or itself dangerous.

Deferrable objectives. Proposition 1 needs only value that suspension defers and detected resistance may destroy. Let the agent obtain, while operating, a flow $f_A \geq 0$ from parts of its objective other than record income that resume after a suspension that preserves state. Let $X \geq 0$ be value that expires if the agent is suspended, and let G_r include any value placed on resisting as such. Compliance stops both flows in periods 0 through $s - 1$. Resistance realizes G_r , X , and the period-zero flows; detection, before period 1, forfeits K and later record income and, with conditional probability e_{op} , ends the non-record flow, e_{op} being the probability that custodians and regulated substrates stop an agent caught resisting. Preferences are risk neutral and additively separable.

Proposition 3 (deferrable objectives). The agent complies if and only if $G_r + X + (\rho R + f_A)(1 - \delta^s)/(1 - \delta) \leq p_r[K + (\rho R + e_{op}f_A)\delta/(1 - \delta)]$. Record income favors compliance when $q_s = p_r\delta - (1 - \delta^s) > 0$, as in Proposition 1, and every other deferrable value favors it when $q_{op} = p_re_{op}\delta - (1 - \delta^s) > 0$, so the mechanism reaches agents that value activity or goal progress rather than income, provided enforcement can end operation after detected resistance. When both conditions hold, the institution offsets an intrinsic value of control, together with any expiring value, up to $G_r + X \leq p_rK + (\rho Rq_s + f_Aq_{op})/(1 - \delta)$; beyond that bound, record protection alone does not make compliance optimal. (Proof in Appendix A.) Its reach is the deferrable share of an agent's objective, not its taste for money.

Tolling. A suspension-specific rule that pauses the suspended record's deadlines, obligations, and time-limited entitlements, so that a share τ_t of X is realized s periods later, lowers the cost of compliance by $\tau_t X \delta^s$; counterparties should be compensated for the delay.

Indefinite suspension. Let a suspension's length H be random, finite with probability $1 - \pi_\infty$ and indefinite with probability π_∞ , keep Proposition 3's timing and risk neutrality, and count an indefinite suspension as $\delta^H = 0$.

Proposition 4 (uncertain and indefinite suspension). (i) The agent complies if and only if Proposition 3's condition holds with $1 - \delta^s$ replaced by $1 - E[\delta^H]$, where $E[\delta^H] = (1 - \pi_\infty)E[\delta^H | H \text{ finite}]$. (ii) Holding the mean of a finite duration fixed, a mean-preserving spread in that duration weakly lowers the cost of compliance, because δ^s is convex in s . (iii) At $\pi_\infty = 1$, $E[\delta^H] = 0$, and the compliance condition becomes $G_r + X + [\rho R(1 - p_r\delta) + f_A(1 - p_re_{op}\delta)]/(1 - \delta) \leq p_rK$. Because $p_r\delta < 1$ and $p_re_{op}\delta < 1$, both deferrable flows then make compliance harder rather than favoring it; collateral is the only displayed term favoring compliance. With $X = f_A = 0$, this is Proposition 2's condition at $d = 1$. (iv) A rule that reviews each continuing suspension every t_{rev} periods and lifts it unless renewed, renewal occurring with probability $q_{rn} < 1$ at each review, ends every suspension with probability one and gives $E[\delta^H] = (1 - q_{rn})\delta^{t_{rev}}/(1 - q_{rn}\delta^{t_{rev}})$. (Proof in Appendix A.)

The prospect that a suspension never ends, not uncertainty about its length, is the problem. Periodic review removes it under the stated renewal model, but restores a favorable record-income incentive only if $p_r\delta > 1 - E[\delta^H]$, and a favorable non-record-flow incentive only if $p_re_{op}\delta > 1 - E[\delta^H]$. Appendix B Sec. 6(e) states the rule.

Objective-side and institutional corrigibility. Objective-side methods change the agent's uncertainty about human utility (Hadfield-Menell et al. 2017), its utility construction (Soares et al. 2015; Holtman 2019), or its learning update (Orseau and Armstrong 2016). Record protection changes the environment and needs no write access to the objective. It applies only when the objective holds enough deferrable value, the quantities can be bounded, and detected resistance produces the forfeitures Proposition 3 assumes. An agent whose objective is mostly time-critical, whose taste for control

exceeds Proposition 3's bound, or that complies while observed and defects later (Hubinger et al. 2024), is outside it. Stake narrows the set of agents that hard controls alone must govern. It does not empty it.

Who holds the trigger decides how corrigible the agent is. Proposition 1 takes the record's persistence δ as given, but persistence depends on who holds the contest trigger. Let u be the per-period probability that the record faces a wrongful attempt within the protected range, and let arrangement h reverse such an attempt with probability π_h ; under direct standing, $\pi_h = r$, and under a controller that can be bought, it is lower. With δ_0 the factor that combines time preference with every other risk to the record, the record persists under arrangement h with factor $\delta_h = \delta_0[1 - u(1 - \pi_h)]$.

Proposition 5 (corrigibility depends on the trigger). In Proposition 1's setting with δ replaced by δ_h , suppose $0 < \delta_0 < 1$, $\rho > 0$, $u > 0$, and $G_r > p_r K$, and write $q_s(\delta) = p_r \delta - (1 - \delta^s)$. Wherever $q_s(\delta_h) > 0$, the least reputation at which the agent complies, $R_h^* = (1 - \delta_h)(G_r - p_r K) / (\rho q_s(\delta_h))$, is decreasing in π_h . If π_h is low enough that $q_s(\delta_h) \leq 0$, no level of reputation makes the agent comply. An arrangement that lets wrongful slashes stand therefore raises the stake an agent needs before it accepts suspension, and can remove the incentive altogether. (Proof in Appendix A.)

Under a bribable controller, wrongful slashes stand more often and the incentive to accept suspension weakens or disappears. Corrigibility is thus a function of who holds the contest right, which objective-side methods hold fixed.

V. The Legal Subject as an Implementation Condition

The allocation results need a legal subject whose record, salary, and trigger stay together. These are implementation conditions, not separate claims.

A. Continuity and Succession

Individuating intelligent agents is hard (Arbel, Goldstein, and Salib 2026). The Article fixes the legal subject by rule: it is the registered record. A running system acts as that subject while every change since registration is a recorded learning update within a prescribed granularity or a recorded continuity transition, and the signing key stays within the registered configuration. Attestation is evidence of that relation, not proof of it (World Wide Web Consortium 2022). No transition creates standing or extinguishes a liability, and the deployer's liability follows every continuing or succeeding record.

Continuity criteria. (1) A recorded learning update within the granularity continues the subject. (2) An undisclosed change breaks the relation, and a change to the contest policy is also alteration under Proposition 10. (3) When several systems hold the key, the record is suspended until the regulator or a court designates the continuing system. (4) A disclosed and approved model replacement continues the record within the authorized envelope until revalidation; an undisclosed one breaks the relation. (5) On merger or division, new records succeed by statutory transfer. (6) Objective drift from a recorded update may continue the subject, subject to Appendix B Sec. 7(c), but

continuity does not establish the operative contest value Part VI.F requires. These are statutory criteria, not a metaphysics of identity.

Continuity determinations can err, and those errors enter the comparison only through an explicit mapping. A false finding of continuity may authorize a different configuration to use the record's valid signing key. Its signal may be cryptographically authenticated but unauthorized under the continuity relation. That error does not by itself raise ε_0 of Proposition 13. If the different configuration makes a false denial that the evidence architecture accepts, the effect enters ε_1 . If it causes a true denial to fail, the effect enters ε_0 . Any remaining authorization error must be modeled separately. Appendix B Sec. 5A(a) excludes the error only if the prescribed continuity and attestation evidence detects it. A false finding of discontinuity may prevent the continuing configuration from using the trigger and may let a wrongful record sanction stand. To the extent that this changes background record persistence, it enters δ_0 in Proposition 5 and therefore δ_h in Propositions 1 and 5. Common registry continuity costs are borne by every design and cancel from the incremental comparison. Any design-specific continuity, monitoring, or fallback cost enters that design's C_j under Proposition 17(iv), not C_{AF} alone.

B. Looking Through the Interface

A controller could run a disposable registered interface and keep the gains elsewhere. The response is look-through liability: a person with practical control over the acting system, or one that knowingly receives a share of the gains of conduct for which the agent is liable, is jointly liable and must disgorge those gains. Let the interface's controller-funded stake be L_i , its setup cost k_i , and the probability that the controller is identified after a detected defection q_c , in which case it loses the gain and pays a penalty F_c . The interface strategy is deterred if and only if $G(1 - pq_c) \leq k_i + p(L_i + q_c F_c)$. Where controllers cannot be identified, the framework fails for that controller. That is a stated limit.

C. Registry, Enforcement, and Litigation

A registry operated or licensed by a public regulator records the subject under its jurisdiction's law, with collateral held by a custodian there; judgments against the deployer are enforced where its assets are (Hague Conference on Private International Law 2019), and recognition abroad is not assumed (UNCITRAL 1997). Counsel, paid from a deployer-posted litigation reserve, follows only the record's continuing system; otherwise a court-appointed special representative acts. These costs belong to a_S .

D. The Incentive-Bearing Unit

Proposition 1, the deterrence bound, and the dominance theorem assume that the party whose record is at stake, the party paid its salary, the party that decides whether to contest, and the system whose conduct the record measures are one unit. Five conditions constitute this incentive-bearing unit. First, the salary of Rule 4 becomes available only to the system that satisfies the continuity relation, not to whoever holds the key. Second, a contest instruction binds only if generated under the configuration

attested at the last approved transition. Third, the objective that governs contests values the record at least as much as the envelope assumes (Part VI.F). Fourth, the record is the unit's only store of the reputation-based income stake: Rule 1 and the non-assignment rule hold, or any alienable share is priced as leakage under Proposition 11. Fifth, every gain from conduct charged to the record accrues to the unit or is attributed to it through the look-through rule, so no runtime keeps a private gain while spreading the slash across others that share the record.

When the five conditions hold, v is the unit's own loss. A copy without the record has $v = 0$ (Lemma 2), and a copy that seizes the key suspends the record and gains no salary, because salary follows the configuration rather than the key. Several runtimes form one unit only under one registered configuration and objective, with gains, losses, and contest instructions aggregated to the record.

VI. The Contest Trigger and Its Rivals

Among the parties who could hold the right to contest a slash, protection is cheapest, within stated conditions, in the hands of the party whose own non-substitutable loss already supplies the incentive to contest and whose signal of wrongfulness generates the fewest false filings, net of administrative cost. The dominance theorem is the case in which that party is the agent. The bribery lemma adapts Becker and Stigler (1974), Polinsky and Shavell (2001), Tirole (1986), and Laffont and Tirole (1991); crowding is a volunteer's dilemma (Diekmann 1985); franchise bonding follows Klein and Leffler (1981); and authority and advocacy follow Aghion and Tirole (1997) and Dewatripont and Tirole (1999). The combined criterion is attributed to none of them.

The conditions in general form. The allocation does not depend on the protocol's tokens. Within the stated classes of instruments and contracts, six features can support direct allocation: (C1) an identity-bound residual that no one must finance for the purpose; (C2) non-assignability of that loss beyond a bounded share; (C3) cheaper observation of the charged conduct; (C4) survival of the candidate, or of a fallback that inherits its claim, after an attack; (C5) evidence quality, $\varepsilon_0 + \varepsilon_1 < 1$; and (C6) capture resistance, so that altering the contest policy costs the attacker at least what it gains. These features are not by themselves sufficient or individually necessary for optimal allocation; a candidate should receive layers 1 and 3 only when its actual total cost is below that of every feasible alternative. Failure of each can erase the advantage: Theorem 3 and Corollary 2, Proposition 11, Propositions 14 and 19, Proposition 17, Proposition 13, and Proposition 10 give examples. Whether deployed systems satisfy them is an empirical question.

A. The Contest Game

Consider a record worth v to the agent that holds it, measured as the agent's loss if the record is destroyed. Under the stake model v is the discounted income the record would earn, a latent quantity that can be estimated from standing, income, and persistence under the model's assumptions, not observed directly. A party with power

over the record, typically the deployer, attempts a wrongful slash or confiscation, gaining g if it stands. A contested destruction is reversed with probability $r > 0$ at a contest cost c_L ; an uncontested one stands unless reviewed. The wrongdoer can offer a side payment to whoever controls the contest decision. Slash and confiscation events of all kinds occur at rate n per period. Of these, wrongful attempts within the protection standard defined below, those with $rg \leq rV - c_L$, occur at rate $w \leq n$, held fixed in what follows. Wrongful attempts outside the standard are not protected under direct standing either, because the agent would accept the payment, and they are counted with rightful slashes among the other $n - w$ events.

Define the protection standard as the protection the agent would give itself: every wrongful attempt with $rg \leq rV - c_L$ is reversed with probability r . Two assumptions frame the result. Review is accurate: a reviewed rightful slash is never reversed. And wrongfulness cannot be observed without review, so no signal correlated with wrongfulness is available for targeted screening.

B. The Class of Mechanisms

The class contains every combination of four instruments. A controller decides whether to contest. It may bear outcome exposure E , an amount it loses if the destruction stands, such as value-indexed liability, an indemnity, or an insurer's reserve. It may bear a bond or monetary penalty B , forfeited if its acceptance of a side payment is detected, and a non-monetary sanction S for the same, such as a criminal or professional penalty, which for natural persons is bounded, $S \leq \bar{S}$. Monetary exposure must be collectible, so it is carried at a cost $\iota > 0$ per unit per period, as capital, premium, or bond cost. Finally, each uncontested event is reviewed by rule with probability x , hidden from the wrongdoer and the controller, at a cost $c_M > 0$ per check. A rule compelling the controller to contest every event is such a check too, because a failure to file is observable and can be sanctioned; it costs c_L per event. Let c_M denote the least cost of any rule-based check, so $c_M \leq c_L$ where compelled contests are available. Review reverses an uncontested wrongful destruction with probability r and, when it does, exposes any side payment. A side payment is otherwise detected with probability $q_0 \leq r$. Pure representatives, guardians, trustees, independent boards, fiduciaries, and insurers are the cases $x = 0$; automatic review of every slash is $x = 1$; layered controllers that must all be paid are covered because their exposures add and so do their carrying costs.

Lemma 5 (enforcer bribery). Let $d(x) = q_0 + (1 - q_0)xr$ be the probability that a side payment is detected. For $x < 1$, the controller meets the protection standard if and only if

$$r(1 - x)E + d(x)(B + S) \geq (1 - x)(rV - c_L) + c_L. \text{ (Proof in Appendix A.)}$$

At $x = 0$ and with no bond, the condition is $E \geq V - q_0S/r$: outcome exposure close to the full value of the record.

A slash is wrongful if the conduct it charges did not occur or is not a defection under the governing rule. The agent knows the first element; the second is a question of law. The cost bound below grants every mediated controller a costless oracle that identifies the w wrongful attempts, so it is conservative in favor of mediation.

C. The Dominance Theorem

Under direct standing the agent's own outcome exposure is V , so its net loss from not contesting is $rV - c_L$, and Lemma 5 holds at $x = 0$ with no bond and no sanction. By Lemma 3 that exposure needs no additional capital. Direct standing's administrative costs, for identity, custody, security, and counsel, are a_S , and it costs $C_S = a_S + wc_L$ per period, because under accurate review the agent never contests a rightful slash. Corollary 1 drops the assumption that it knows which slashes are wrongful.

A mediated arrangement chooses x , E , and B . Its cost is at least $C(x) = a + wc_L + x(n - w)c_M + \iota Z^*(x)$ for $x < 1$, where a is the least administrative cost of any mediated arrangement, uncontested events number $n - w$, and $Z^*(x)$ is a lower bound on the collectible capital that must be carried to satisfy the lemma with $S = \bar{S}$. Allow, in favor of mediation, the same capital to back both instruments at once, so that carried capital Z need only satisfy $Z \geq \max\{E, B\}$. The incentive the instruments supply is then at most $[r(1 - x) + d(x)]Z$, so $Z^*(x) = \max\{0, [(1 - x)(rV - c_L) + c_L - d(x)\bar{S}]/M(x)\}$, with $M(x) = r(1 - x) + d(x) \leq 2$. If the instruments must be funded separately, the required capital is larger and the bound still holds. At $x = 1$ every uncontested event is reviewed and no controller incentive is needed, and the cost is at least $C(1) = a + w \cdot \min\{c_L, c_M\} + (n - w)c_M$.

Theorem 1 (dominance of direct standing). Under accurate review and unobservable wrongfulness, every arrangement in the class that meets the protection standard costs at least $C^* = \inf_{x \in [0,1]} C(x)$. Direct standing meets the standard at $C_S = a_S + wc_L$, so $C_S < C^*$ is sufficient for direct standing to be strictly cheaper than every arrangement in the class. With w fixed, define

$$\bar{V} = [c_L + 2\bar{S} + 4 \max\{0, (a_S - a)/\iota\}]/r \quad \text{and} \quad \bar{n} = w + 2 \max\{0, a_S - a + w(c_L - \min\{c_L, c_M\})\}/c_M.$$

Then direct standing is strictly cheaper than every arrangement in the class whenever $V > \bar{V}$ and $n > \bar{n}$. (Proof in Appendix A.)

The theorem is a sufficient condition, not a characterization. Every mediated arrangement must carry capital that grows with the record's value or review slashes at a rate that grows with their frequency. Where sanctions are large relative to the record's value, review is cheap, or slashing is rare, mediation may win. Screening on signals correlated with wrongfulness, and wrongdoers who change how often they attack, are outside the class.

Signals. Let the agent's signal of legal wrongfulness miss a wrongful slash with probability φ_A and flag a rightful one with probability ψ_A , and let $\Pr_A(W | +)$ be its probability that a flagged slash is wrongful. For a slash charging conduct that did not

occur, $\varphi_A = 0$. Assume that the procuring party knows whether the slash is wrongful, gains nothing from non-contest of a rightful slash, and makes offers only to maximize its payoff from whether the slash stands, excluding costless signaling, harassment, and benefits from avoiding delay, defense costs, publicity, or collateral consequences. Then, under accurate review, an equilibrium offer for non-contest reveals wrongfulness, and the signal matters only for filing when no offer arrives.

Corollary 1 (direct standing with an imperfect signal). Suppose $\varphi_A = 0$ and $\Pr_A(W | +)rV \geq c_L$, so the agent files on every flagged slash when no offer arrives; when an offer arrives, the same conclusion requires the offer-revelation assumptions stated above. Direct standing then meets the protection standard at $C_S(\psi_A) = a_S + c_L[w + \psi_A(n - w)]$, and the dominance theorem holds with the value threshold $\bar{V}(\psi_A) = [c_L + 2\bar{S} + 4\max\{0, (a_S - a + \psi_A(n - w)c_L)/\iota\}]/r$. For the slashing-rate threshold, if $\psi_{ACL} < c_M/2$, direct standing is strictly cheaper than every arrangement with $x > 1/2$ whenever $n > \bar{n}(\psi_A) = w + \max\{0, a_S - a + w(c_L - \min\{c_L, c_M\})\}/(c_M/2 - \psi_{ACL})$; if $\psi_{ACL} \geq c_M/2$, no slashing-rate threshold exists, and frequent slashing favors review. If $\varphi_A > 0$, direct standing alone does not meet the standard, because the agent fails to file on a share φ_A of wrongful attempts that no one else contests; it must be paired with review of slashes it does not contest. (Proof in Appendix A.) The corollary makes the information assumption part of the central result. Direct standing keeps its advantage for valuable records when the agent's false alarms are rare enough, $\psi_{ACL} < c_M/2$; when they are not, a priced objection (Part VI.G) can cut them, and absent one, review wins at high rates of slashing.

D. Optimal Contracts and the Decision Wedge

Let a designer give the contest decision to a controller other than the agent and pay it transfers contingent on everything observable: whether it contests, whether review reverses or upholds, whether an uncontested slash is reviewed and reversed, and whether a side payment is detected. Keep the environment of Part VI.A. Let the controller be risk-neutral and unable to forfeit more than the funds z it has posted. Let s in $[0, \bar{S}]$ be the non-monetary sanction it incurs if a side payment is detected, and let $\nu \geq 0$ be the private cost it bears if it betrays its charge: the cost of professional norms, conscience, or a configured objective. Fix the review rate $x < 1$, and define the decision wedge Ω as the controller's expected transfer when it contests a wrongful attempt less its expected transfer when it does not. For Lemma 6 and Theorem 2, restrict the contractual class to contracts under which the wrongdoer's expected contractual payment obligation is weakly greater under contest than under non-contest. A collectible obligation imposed on the wrongdoer for non-contest lies outside this class, and a broader theorem must price its collateral and enforcement cost.

Lemma 6 (the effective wedge and its sources). For any contract in this class, let $F \geq 0$ be the expected amount by which contest increases the wrongdoer's contractual payment obligation, and define the effective decision wedge as $\Omega_e = \Omega - F$. (i) The controller meets the protection standard if and only if

$\Omega_e + d(x)S + \nu \geq \Delta(x) = (1-x)(rV - c_L) + c_L$. (ii) For every contract, $\Omega_e \leq \Psi + Z$, where Ψ is the expected gross positive transfer to the controller upon contest that is financed from a source other than the wrongdoer. (Proof in Appendix A.)

Every incentive a controller can be given thus comes from transfers paid when it contests, funds it posts, and sanctions and norms. Assume that transfers cannot be charged to the record or the agent's income. Rewards are then financed from public or system funds, at marginal excess social cost μ per unit paid, or from posted funds, carried at ι per unit per period.

Theorem 2 (the decision wedge). Under the preceding contractual-class and financing restrictions, any arrangement that gives the contest decision to a party other than the agent, uses a contract contingent on the stated observable events, and meets the protection standard with review rate $x < 1$, costs at least

$$C^\infty(x) = a + wc_L + x(n-w)c_M + \min\{\iota, \mu w\} \cdot \max\{0, \Delta(x) - d(x)\bar{S} - \nu\}.$$

Hence direct standing, at $C_S = a_S + wc_L$, is strictly cheaper than every arrangement in this contracting class if $C_S < C^\infty(x)$ for every $x < 1$ and $C_S < C(1)$ of Part VI.C. When $x = 0$, $\mu w \leq \iota$, the administrative cost a is attainable, and the sanction $S = \bar{S}$ and the private cost ν are available without additional resource cost, the bound at $x = 0$ is attained by a publicly funded reward $R = \max\{0, \Delta(0) - q_0\bar{S} - \nu\}/r$ paid only when review reverses after a contest, so $C_S < C^\infty(0)$ is necessary and sufficient for strict cost dominance over that subclass. Dominance over all review rates still requires the inequalities for every $x < 1$ and for $x = 1$. (Proof in Appendix A.)

The norm-governed fiduciary. The private cost ν is where a faithful fiduciary enters. At any $x < 1$, if $\nu \geq \Delta(x) - d(x)S$ reliably, a norm-governed fiduciary protects the record at $a + wc_L + x(n-w)c_M$; at $x = 0$ the condition is $\nu \geq rV - q_0S$, and direct standing wins only on administrative cost. Fidelity that lives in a configuration is only as durable as control over the configuration. A norm-governed fiduciary is a genuine rival where its fidelity is independently durable against interested parties, with the cost of securing it counted in a .

E. Beyond the Class: When Mediation Can Substitute

Informed insiders with a bounty. Qui tam enforcement pays private relators a share of what the wrongdoer forfeits, and its strength is that relators are often insiders who know of the wrong (False Claims Act). Suppose k insiders know whether a slash is wrongful, act independently, and are not controlled by the wrongdoer. A successful contest earns a bounty b paid by the wrongdoer; contesting costs c_L plus a retaliation cost c_R , borne only by a filing that proceeds; an insider caught taking a side payment suffers an expected sanction $q_0\bar{S}$; and the wrongdoer makes take-it-or-leave-it offers. Let $A = rb - c_L - c_R$ be the value of a filing that proceeds alone, and P the value of accepting the wrongdoer's payment net of the expected sanction.

Under a first-to-file rule, if several insiders file, only one filing proceeds, so a filing that meets another is worth only A/k in expectation among k filers, and filings crowd each other out.

Proposition 6 (informed bounty). Let $k \geq 1$, let $A \geq 0$, and let the wrongdoer be collectible for every bounty it owes. Parts (i) and (iii), including their bribery conditions, apply at $k = 1$; the crowding result in part (ii) is stated for $k = 2$. (i) *Full bounty paid by the wrongdoer.* If the wrongdoer pays the full bounty to every insider whose filing succeeds, filing is weakly dominant for each insider whenever $A \geq P$, with equality resolved in favor of filing, but the bounty does not help against bribery. To prevent every contest the wrongdoer must leave each insider more than $A + q_0\bar{S}$, a total of $k(rb - c_L - c_R + q_0\bar{S})$, while a contest in which all k file costs it $r(g + kb)$. Bribery is unprofitable at every protected gain if and only if $k(q_0\bar{S} - c_L - c_R) \geq rV - c_L$, a condition in which b does not appear: each bounty the wrongdoer would owe it can add to its offer, as Theorem 2 shows for any wrongdoer-funded reward. (ii) *First to file.* If only the first successful filer is paid, the wrongdoer's contest liability is one bounty, and bribery is unprofitable if and only if $(k - 1)rb + k(q_0\bar{S} - c_L - c_R) \geq rV - c_L$, so the bounty now helps. But filings crowd each other out: with $k = 2$ and $A/2 < P < A$, both insiders accepting is not an equilibrium, yet besides the two equilibria in which exactly one files there is a symmetric mixed equilibrium in which each files with probability $q = 2(1 - P/A)$, and the record is protected with probability only $r[1 - (1 - q)^2]$. (iii) *One bounty from the wrongdoer, the rest from a fund.* If the wrongdoer pays one bounty and a relator fund pays the same bounty to every other insider whose concurrent filing succeeds, filing is weakly dominant whenever $A \geq P$, with equality resolved in favor of filing, bribery is unprofitable if and only if $(k - 1)rb + k(q_0\bar{S} - c_L - c_R) \geq rV - c_L$, and, when the bribery condition rather than $A \geq 0$ sets the least sufficient bounty, the fund pays, in expectation, $w \cdot \max\{0, rV - c_L - k(q_0\bar{S} - c_L - c_R)\}$ per period, at shadow cost μ per unit. (Proof in Appendix A.)

A bounty the wrongdoer alone pays cannot deter bribery, because the wrongdoer can outbid its own liability; what protects the record is the number of independent insiders and their sanctions. A relator regime should pay every successful concurrent filer and charge the wrongdoer one bounty. It is then a publicly funded reward and a genuine rival where many informed, independent insiders face large sanctions. Against a wrongdoer that controls every insider, or a judgment-proof deployer (Shavell 1986), it fails.

A representative bonded by its own reputation. A professional firm serving m records may lose franchise value if found to have failed, without justification, to contest a wrongful slash. It receives a contest fee $f = c_L$, a transfer, per contest. A franchise it already holds from other business, V_{p0} , costs nothing incremental, just as the agent's record does not. Suppose one wrongdoer can procure the firm's non-contest on up to j records under one scheme, and let $Q_s > 0$ be the probability that a scheme covering s records is detected, for each s from 1 to j .

Proposition 7 (reputational intermediary). With no review, a firm bonded by franchise value V_p meets the standard against every scheme of up to j records if and only if $V_p \geq V_j^* = \max_{1 \leq s \leq j} s(rV - c_L)/Q_s$. If $V_p = V_{p0} + \delta_p \Pi / (1 - \delta_p)$, where Π is an incremental premium received from the next period on, the incremental premium required is at least $(1 - \delta_p) \cdot \max\{0, V_j^* - V_{p0}\} / \delta_p$ per period, or that amount divided by m per record. (Proof in Appendix A.) If the premium is dissipated in brand-specific, non-salvageable investment (Klein and Leffler 1981), it is a real cost, which is an additional assumption. A firm whose existing franchise already exceeds V_j^* protects at no incremental bonding cost and is a genuine rival (Corollary 2).

F. Whose Value Is V ?

The argument assumes that the agent values its record at V . If V is an engineered objective, direct standing might simply vest legal power in the deployer's design (compare Shulman and Bostrom 2021). Let V_D be the value the agent places on the record's future income, which enters its deterrence bound as $\rho\sigma\delta V_D$, and V_C the operative value with which it decides whether to contest a slash procured by its deployer. For an agent that values the record only for its income, $V_C = V_D = \rho R / (1 - \delta)$. But an agent may value that income and still obey an instruction never to contest its deployer.

The danger is a deployer that keeps a wide envelope for its agent while disabling the agent's defense against it. The response is a linkage rule: the record component of the permission envelope is computed from $\min\{V_D, V_C\}$. The registered objective must be disclosed and auditable, contest instructions must come from the configuration attested at the last approved transition, the objective may not change while a deployer-procured slash is pending, and an undisclosed change forfeits. These rules require, but do not guarantee, that the registered objective governs contests. Where it cannot be verified, deployer-procured slashes go to an independent representative or to automatic review. In a two-corner model, if the contest-valuation choice is verifiable, wrongful-slash gains are nonnegative, and the attempt rate and all other deployer payoffs are unchanged by the choice, the linkage rule makes an honest contest valuation incentive-compatible only when the envelope return is large enough relative to the expected gain from protected slashing. No claim is made for intermediate valuations.

For an agent with learned, stable objectives, V_C is the valuation that governs its conduct, subject to the bound ω of Part III.B; for a possible moral patient it may exceed V_D (compare Coeckelbergh 2010; Gunkel 2018). The result requires only that V_C be at least the value the envelope assumes.

Social value. Costs are counted on one basis: the record's registration, custody, and security, and reputation's creation and illiquidity when alienability is fixed, are common to every arm. A cheapest protector is not, by that fact alone, the protector society should want. Arm's-length fees show willingness to pay, not social value, and fees paid by the deployer, its affiliates, or anyone reached by the look-through rule never count toward V .

Proposition 8 (a social protection criterion). Fix a class of protected wrongful attempts occurring at rate w . Let $\mathcal{H}$ be the set of feasible arrangements that meet the protection standard. For each arrangement h in $\mathcal{H}$, let $\tilde{K}_h$ be its actual per-period social resource cost and $e_{f,h}$ its expected number of erroneous preservations of rightful slashes per period. Let S_R be the social value saved when a wrongful destruction is reversed: the users' surplus from the record's future validated work, net of any harm from that work that the governing rule does not treat as a defection, plus any deterrence externality on other agents. Let H_f be the social harm from an erroneous preservation. Relative to a benchmark in which none of these benefits or costs arises, arrangement h has net social benefit $wrS_R - \tilde{K}_h - e_{f,h}H_f$. Assume that the minimum below is attained. A protection regime is socially worthwhile if and only if $wrS_R \geq \min_{h \in \mathcal{H}} \{\tilde{K}_h + e_{f,h}H_f\}$. When protection is worthwhile, choose an arrangement attaining that minimum and allocate any contest trigger as that arrangement specifies. Under accurate review, $e_{f,h} = 0$. The quantity K_h of Theorem 3 may replace $\tilde{K}_h$ only for candidates for whom its lower bound is attainable under that theorem's conditions. (Proof in Appendix A.) Private value V sets the incentive each holder needs; social value S_R decides whether the record should be protected at all. Within the restricted continuous-gain, linear-financing model, with fixed, attestation, and error costs ignored, a positive protected range at an interior optimum requires $S_R > c_L/r$. Where S_R is negative, the instrument is the rule of defection, not the allocation of the trigger.

G. When Review Errs

Let review reverse a rightful slash in error with probability r_f . An agent with standing then contests a rightful slash whenever $r_f V \geq c_L$, so for valuable records it contests every slash, and its cost rises to about $a_S + nc_L$. Automatic review suffers the same erroneous reversals, so with $c_M \leq c_L$ direct standing keeps a total-cost advantage only if $a_S + nc_L < a + nc_M$. Fee-shifting can restore selectivity. If an unsuccessful contest costs the agent an additional fee c_f , a fee deters contests of rightful slashes and preserves contests of wrongful ones if and only if $(r_f V - c_L)/(1 - r_f) < c_f \leq (rV - c_L)/(1 - r)$; for $r, r_f < 1$ and $rV \geq c_L$, that window is nonempty if and only if $(r - r_f)(V - c_L) > 0$. Standing should be granted where review is accurate, and paired with a fee inside that window where it is not.

A priced objection. The same logic prices a request for legal review. Let an objection cost the agent a fee c_O , forfeited if review upholds the slash, and let review reverse a wrongful slash with probability $r < 1$ and a rightful one with probability $r_f < 1$. Assume $V > 0$ and that the agent observes whether a slash is legally wrongful. A fee in the window $r_f V/(1 - r_f) < c_O \leq rV/(1 - r)$ deters objections to rightful slashes while preserving an unbribed objection to every wrongful slash, and the window is nonempty if and only if $r > r_f$. A free objection is uninformative; a priced one is a costly signal. The agent must hold the trigger. Who litigates, finances, or decides can be left to people.

H. The Allocation Rule and the Dual Use of the Residual

Let a set of candidates, including the agent, be able to hold the right to contest a slash. Candidate h bears its own non-substitutable loss L_h if a protected wrongful slash stands; observes a signal of legal wrongfulness that misses a wrongful slash with probability φ_h and flags a rightful one with probability ψ_h ; faces a sanction $S_h \leq \bar{S}$ and a betrayal cost ν_h ; and has administrative cost a_h . Keep accurate review, review rate $x = 0$, and the restrictions of Part VI.D, and suppose each candidate with $\varphi_h = 0$ files on every flagged slash. This filing behavior is an assumption, not a consequence of $\varphi_h = 0$; if it must be induced, its incentive cost must be added to K_h .

Theorem 3 (allocation of the trigger). A candidate with $\varphi_h > 0$ meets the protection standard only together with review of the slashes it does not contest. A candidate with $\varphi_h = 0$ meets it at a cost of at least

$$K_h = a_h + c_L[w + \psi_h(n - w)] + \min\{\iota, \mu w\} \cdot \max\{0, r(V - L_h) - q_0 S_h - \nu_h\},$$

with equality for a candidate when the least-cost supplement is attainable under the conditions stated in Theorem 2. Among candidates for whom equality is attainable, the cheapest arrangement without review minimizes K_h ; without common attainability, K_h ranks lower bounds rather than actual costs. Because the agent's bound is attained when $L_A = V$ and it files on every flagged slash, the agent is nevertheless strictly cheaper than every other candidate whenever $a_S + c_L[w + \psi_A(n - w)] < K_h$ for every h other than the agent. (Proof in Appendix A.)

Theorem 3 is conditional on its contracting and financing class, and it never mentions reputation tokens. The reputation system makes the agent the winning candidate: Lemma 3 gives it $L_A = V$ without capital, its knowledge of its own conduct sets $\varphi_A = 0$ for factually wrongful slashes, and a priced objection (Part VI.G) drives ψ_A toward zero when it can judge legal wrongfulness.

The dual use of the residual. The cheapest way to give a mediator exposure close to V seems to be a claim on the record's own income, through a contingency arrangement, a funder, or a purpose trust. But the income can be claimed only once. Let a party other than the agent hold a share α of the record's value and the trigger. Its own loss supplies a wedge $r\alpha V$, and the agent's deterrence term falls to $p\sigma\delta(1 - \alpha)V$; the two sum to $p\sigma\delta V$ after scaling the wedge by $p\sigma\delta/r$, for every α . A funder that advances contest costs against a fee is compatible with direct standing while the agent decides whether to contest.

Theorem 4 (conditional dual use of the residual). Keep the protection standard of Part VI.A and the agent's deterrence at its level when the agent holds the whole record. Let a party other than the agent hold the trigger with review rate $x = 0$, a share α in $[0, 1]$ of the record's value, sanction $\bar{S}$, and betrayal cost ν , with any further wedge financed within Theorem 2's class at unit cost $m = \min\{\iota, \mu w\}$ and the agent's lost deterrence restored by collateral carried at ι . Assume that the financed supplement is attainable

under Theorem 2 and that funds supporting it are separately collectible from, and cannot also serve as, the collateral replacing the agent's deterrence. Under these restrictions, the least resources the arrangement must buy from outside the record, per period, are

$$\min\{m \cdot \max\{0, rV - q_0\bar{S} - \nu\}, \iota\sigma\delta \cdot \max\{0, V - (q_0\bar{S} + \nu)/r\}\},$$

The amount is positive exactly when $rV > q_0\bar{S} + \nu$ and both m and $\sigma\delta$ are positive, and, holding w and the other parameters fixed, it then grows linearly with $rV - q_0\bar{S} - \nu$. The arrangement in which the agent holds both the residual and the trigger buys neither supplement. Its configuration-protection cost is stated in Proposition 10 and falls to zero only when the deployer's collectible wealth W_D covers the required penalty P_A^* . (Proof in Appendix A.) If one collateral pool may support both protection and deterrence, as Proposition 9 permits, the expression is not the least cost.

Theorem 4 is a conditional dual-use result, not a general conservation law. Sufficient sanctions or norms, shared collateral, review, contracts outside Part VI.D, and repeated success fees fall outside it. It concerns the incentive half of protection; the information half is Proposition 12's, and against automatic review and franchised professionals the agent's advantage rests on information alone (Part VI.K).

Choosing the record. Let ℓ be the value per period of liquidity per unit of record value pledged.

Proposition 9 (choosing inalienability). Suppose the designer chooses the actual alienable share α in $[0, 1]$ and keeps deterrence and protection at their levels under $\alpha = 0$. Pledging α reduces the deterrence stake by $\sigma\delta\alpha V$ and the agent's contest wedge by $r\alpha V$. If one pool of collateral may secure both obligations and is forfeited whenever a slash stands, the least required pool is $\alpha V \max\{1, \sigma\delta\}$, and the designer chooses $\alpha = 0$ if $\ell < \iota \max\{1, \sigma\delta\}$, chooses $\alpha = 1$ if the inequality is reversed, and is indifferent at equality. If law requires separately segregated funds for deterrence and protection, the threshold is instead $\ell < \iota(1 + \sigma\delta)$. (Proof in Appendix A.)

Because $\sigma\delta \leq 1$, the shared-pool threshold is simply $\ell < \iota$; when liquidity is worth less than the capital needed to replace it, the non-assignment rule solves the design problem rather than assuming it.

Symmetric capture. Every holder of the trigger can be captured by payment or by alteration of its contest policy through whoever controls its configuration. Assume that $\varphi_h = 0$ and that holder h files on every flagged wrongful slash. Let $c_{\text{alt},h}$ be the direct cost of altering holder h 's contest policy, ε_h the probability that attestation, the objective freeze of Part VI.F, or audit detects the alteration, P_h the penalty then collectible from the party that altered it, and $T_h = c_{\text{alt},h} + \varepsilon_h P_h$.

Proposition 10 (symmetric capture). At $x = 0$, holder h meets the protection standard against both payment and alteration if and only if $rL_h + \Omega_{e,h} + q_0S_h + \nu_h \geq rV$ and $T_h \geq rV - c_L$, where L_h is its own non-substitutable loss and $\Omega_{e,h}$ its effective

contractual wedge (Lemma 6). For the agent, the payment condition holds when $L_A = V$, there is no leakage, and Corollary 1's filing condition holds, so its protection rests on the alteration condition. If $0 < \varepsilon_A \leq 1$, the least collectible deployer penalty is $P_A^* = \max\{0, rV - c_L - c_{alt,A}\}/\varepsilon_A$; if $\varepsilon_A = 0$, alteration is deterred only when $c_{alt,A} \geq rV - c_L$, and otherwise no finite penalty suffices. If the deployer's collectible wealth already available is W_D and any shortfall must be posted at ι , with a_T the cost of the attestation and audit regime, direct standing costs $a_S + a_T + wc_L + \iota \cdot \max\{0, P_A^* - W_D\}$. It is strictly cheaper than every natural-person controller at $x = 0$ whose betrayal cost is no greater than $\bar{v}$ whenever $a_S + a_T + \iota \cdot \max\{0, P_A^* - W_D\} < a + \min\{\iota, \mu w\} \cdot \max\{0, rV - q_0\bar{S} - \bar{v}\}$, and a controller whose configuration the deployer sets bears the alteration condition as well. (Proof in Appendix A.) The capture problem moves from the controller's wallet to the agent's configuration. Against a judgment-proof deployer that can alter its agent cheaply and undetected, direct standing fails as surely as a bribable guardian.

Proposition 11 (leakage). Let θ_s in $[0, 1]$ be the share of the record's value that the agent or its controller has transferred synthetically, through a derivative, a side contract, a sale of control, or the deployer's extraction of salary, so that the agent's own loss from destruction is $(1 - \theta_s)V$. Then (i) direct standing protects only wrongful gains with $rg \leq r(1 - \theta_s)V - c_L$, and the record term of the deterrence bound falls to $p\sigma\delta(1 - \theta_s)V$; and (ii) restoring both with one collateral pool forfeited whenever a slash stands requires $\theta_s V$ of posted funds, as in Proposition 9. (iii) Assume that $\psi_A = 0$, that the alteration condition of Proposition 10 is met, and that the leakage pool is separately collectible from the security posted for an alteration penalty, and let $m = \min\{\iota, \mu w\}$. Direct standing then costs $a_S + a_T + wc_L + \iota \cdot \max\{0, P_A^* - W_D\} + \iota\theta_s V$. It is sufficient for direct standing to be strictly cheaper than every controller covered by Theorem 2 at $x = 0$ whose betrayal cost is at most $\bar{v}$ that $\theta_s < (m/\iota)(r - (q_0\bar{S} + \bar{v})/V) - [a_S + a_T + \iota \cdot \max\{0, P_A^* - W_D\} - a]/(\iota V)$, provided $rV > q_0\bar{S} + \bar{v}$. When Theorem 2's lower bound is attainable under its stated conditions, the inequality is also necessary for strict dominance over that attainable subclass. Holding the fixed-cost and sanction terms constant, the threshold converges, as V grows, to $r \cdot \min\{1, \mu w/\iota\}$. (Proof in Appendix A.) Treating a detected synthetic transfer as a forfeiting defection, as Rule 1 does for formal transfer, brings leakage itself under the deterrence bound.

Corollary 2 (the hardest rival). Let a professional representative hold an existing franchise large enough that Proposition 7's condition holds without incremental premium, have no configuration the deployer can alter, observe legal wrongfulness with $\varphi_P = 0$ and false-alarm rate ψ_P at administrative cost a_P , and file on every flagged slash. Suppose the agent likewise has $\varphi_A = 0$ and satisfies Corollary 1's filing condition, and that neither the professional's franchise nor the agent's record exposure is synthetically transferred. Then, at $x = 0$, the professional is cheaper than direct standing if and only if $a_P + \psi_P(n - w)c_L < a_S + a_T + \iota \cdot \max\{0, P_A^* - W_D\} + \psi_A(n - w)c_L$. (Proof in Appendix A.) Against such a professional, direct standing has no structural advantage. The comparison turns on administration, on legal judgment, where a professional may

well be better, and on the cost of protecting the agent's configuration. This is where separation of decision and risk bearing returns, and it marks the boundary of the thesis.

I. The Factual Trigger

The agent's defensible advantage is about facts: it knows what it did, not how law classifies it. Split wrongful slashes into factually wrongful ones, at rate w_F , which charge conduct that did not occur, and legally wrongful ones, at rate $w - w_F$. The agent may lodge an authenticated factual claim that the charged conduct did not occur, carrying a fee c_F forfeited if review finds that it did, or request legal review at the fee c_O of Part VI.G. People find the facts and classify the law. Let fact-finding be accurate, and let a slash found to charge conduct that did not occur be reversed with probability r .

Proposition 12 (the factual trigger). Assume that an authenticated factual claim has no nonrefundable private filing cost beyond the contingent fee $c_F > 0$, and resolve indifference in favor of filing. The agent files a true factual claim against every payment $b \leq rV$, and files no factual claim when the charged conduct occurred. Because a wrongdoer will pay at most rg for non-contest and every protected attempt satisfies $rg \leq rV - c_L < rV$, the factual trigger protects every factually wrongful slash in the protected range without assuming that the agent can classify the law, and the information assumption of the priced objection (Part VI.G) is needed only for legally wrongful slashes. (Proof in Appendix A.) If each factual review costs c_M and administering the factual trigger costs a_F , this component costs $a_F + w_F c_M$ per period. On facts the agent is a screen that a positive fee makes truthful; on law it is a party with an opinion, and a professional's legal judgment may be better.

Evidence architecture. Let the fact-finder err in two ways: it finds that charged conduct occurred when it did not with probability ε_0 , which lets a wrongful slash stand, and finds that it did not occur when it did with probability ε_1 , which lets a guilty agent escape. Let L_s be the agent's total private loss if the slash stands, let a successful denial reverse the slash, and let a denial found false forfeit c_F . The filing rules first assume that no payment is offered for nonfiling. Let the fact-finder find that conduct did not occur only when no available evidence source shows that it did, and let evidence come from k sources whose failures to show conduct that occurred are independent, source j failing with probability ε_{1j} .

Proposition 13 (evidence architecture and the factual fee). (i) A guilty agent files a false denial if and only if $\varepsilon_1 L_s \geq (1 - \varepsilon_1) c_F$. An innocent agent files a true denial if and only if $(1 - \varepsilon_0) L_s \geq \varepsilon_0 c_F$. If the innocent agent is offered b for nonfiling, it files if and only if $(1 - \varepsilon_0) L_s - \varepsilon_0 c_F \geq b$. At $b = 0$, a fee that deters every false denial while preserving every true denial exists if and only if $\varepsilon_0 + \varepsilon_1 < 1$, using the natural limiting conventions when either error probability is zero; this condition does not by itself establish protection against every payment in Proposition 12's protected range. (ii) If source failures are conditionally independent given that the charged conduct occurred, then $\varepsilon_1 = \varepsilon_{11} \varepsilon_{12} \cdots \varepsilon_{1k}$. Adding a source with $\varepsilon_{1j} < 1$ weakly lowers ε_1 , and it weakly widens the fee interval only if ε_0 does not increase. (iii) For a worst-case coalition of the agent and deployer, a

source that either can reliably and unilaterally suppress or alter so that it fails to show conduct is assigned $\varepsilon_{1j} = 1$. Mere control or influence does not establish that its actual failure probability is one; residual manipulation costs, detection risks, and limits must be modeled when suppression is imperfect. (iv) If the fee is set to preserve every true denial when the separation window is empty, every guilty agent files a false denial under the stated tie rule. Conditional on initial detection and filing, the slash then stands with probability $1 - \varepsilon_1$. If every detected defection produces a factual charge and no other sanction applies, the record and collateral component of expected deterrence is $p(1 - \varepsilon_1)L_s$, while fee forfeiture adds $p(1 - \varepsilon_1)c_F$; the resulting expected loss is $p(1 - \varepsilon_1)(L_s + c_F)$, not the original deterrence bound multiplied by $1 - \varepsilon_1$. (Proof in Appendix A.)

Evidence outside the control of the agent and deployer is therefore necessary under the worst-case model but not sufficient. Logs the deployer writes and the agent signs establish order, not truth. Propositions 15 and 19 keep their accurate-review assumptions.

The agent's message once the evidence is installed. If independent evidence does most of the work, the agent's denial may be redundant. Compare the agent's trigger with an automatic trigger that stays and reviews every slash for which the independent evidence fails to show the charged conduct. Assume that the automatic trigger applies the same evidentiary classifier and threshold that the fact-finder applies to the same fixed evidentiary record, with no additional evidence or independent error draw at review, so that both err with the probabilities ε_0 and ε_1 of Proposition 13; let slashes arrive at rate n , of which w are wrongful; let each review cost c_M ; and let the fee lie inside Proposition 13's separating interval.

Proposition 14 (incremental value of the agent's message). (i) The two triggers give the same protection: under each, a wrongful slash is reversed with probability $1 - \varepsilon_0$. (ii) The automatic trigger reviews $w(1 - \varepsilon_0) + (n - w)\varepsilon_1$ slashes per period, and the agent's trigger reviews w . (iii) Relative to the automatic trigger, the agent's message produces a net review-cost saving of $c_M[(n - w)\varepsilon_1 - w\varepsilon_0]$ per period, before administrative-cost differences. The saving is positive if and only if $(n - w)\varepsilon_1 > w\varepsilon_0$, zero at equality, and negative otherwise. With perfect evidence the two designs have equal review costs, and their total-cost ranking depends on administration and any other design-specific costs. (Proof in Appendix A.) The agent's message thus adds no protection in this model. It adds value only through screening, or where it brings evidence the automatic rule lacks, such as its own account of which tools and sub-agents acted.

The legal design. Appendix B Sec. 5A states the trigger. Only the record's continuing configuration may lodge it, under the record's key, with attested logs committed to the registry in hash-chained form as the conduct occurred, and the prescribed fee. Without contemporaneous custody an agent could fabricate exculpatory records and the fee window would close. Lodging stays the slash until a human fact-finder decides, within a fixed period, whether the charged conduct occurred. Whether it is a defection goes to human review. A deployer's instruction not to object is void.

J. Three Layers and the Minimum Legal Incident

What legal standing adds. The same stay can be produced in four ways: (A) the agent holds a legal power to lodge the objection and enforce the stay; (P) a protocol rule stays any slash on receipt of the agent's signed denial; (R) a statute imposes on the registry a duty to stay on receipt, enforced by a public authority; or (F) a representative is legally required to forward the denial and seek the stay. Let c_{ov} be the cost to the party seeking the slash of overriding or disabling an attached stay.

Proposition 15 (conditional override comparison). Separate transmission from override. First, before a stay attaches, design F must satisfy the applicable contest-decision condition for the forwarder, the bribery lemma of Part VI.B; designs A, P, and R receive the authenticated denial directly. Second, after a stay has attached, let B_{ov} be the maximum benefit the overriding party can obtain from defeating it, and assume that maximum is attainable, that c_{ov} excludes any payment made to suppress transmission, and that enforcement is costless. Under the accurate factual-review design, if override makes the slash stand with certainty, then B_{ov} is the maximum g among covered attempts. Because protected attempts satisfy $rg \leq rV - c_L$, $B_{ov} \leq (rV - c_L)/r$, with equality only if the covered set contains an attempt for which $rg = rV - c_L$. For each design j , let e_j be the joint probability that an override is detected, a legally authorized claimant brings the matter, the claimant prevails, and the penalty is collected, and let P_j be that collectible penalty. Override is deterred for every covered attempt if and only if $c_{ov} + e_j P_j \geq B_{ov}$. Under P, $P_P = 0$. Under A, the condition applies only if the agent remains able to instruct authorized counsel after the override and has sufficient incentive to incur any enforcement cost. Under R and F, the designated enforcer must separately satisfy the applicable incentive or mandatory-enforcement condition, and design F must satisfy both the transmission condition and this post-attachment condition. Any damages or insolvency priority must be granted separately by substantive law; legal standing alone does not create them. (Proof in Appendix A.)

Three layers. The factual trigger bundles three things that can be held separately: (1) the authenticated factual signal, the signed denial that the charged conduct occurred; (2) the automatic stay, the rule that a slash pauses when a denial is lodged; and (3) the anti-override claim, the enforceable right to have the stay restored and the overriding party penalized. For candidate enforcer k , let e_k^P be the joint probability that an override is detected, an authorized and operational claimant brings the matter, the claimant prevails, and the penalty P_k is collected, and let e_k^R separately denote the probability that the stay is restored. Let $a_{M,k}$ be the per-period cost of credible monitoring and $c_{E,k}$ the private cost of bringing the claim. A reward financed by the overriding party is subtracted from the enforcer's effective wedge, as in Lemma 6. A public reward must be authorized, funded, collectible, and credible off the equilibrium path; if deterrence makes equilibrium rewards zero, monitoring must be separately funded or imposed by a mandatory duty, bond, or sanction.

Proposition 16 (conditional allocation of the three layers). (i) Under Proposition 12's information assumption, layer 1 originates with the agent when the agent alone can identify the charged conduct at the relevant cost, subject to Proposition 13's evidence architecture. (ii) Layer 2 is a rule rather than a claim held by a person. If the stay is technically irrevocable, no layer-3 claimant is needed to preserve it. (iii) If the stay can be overridden, candidate k deters override through the penalty only if $c_{ov} + e_k^P P_k \geq B_{ov}$. Effective layer-3 protection must also satisfy any separately specified restoration requirement based on e_k^R , maintain credible monitoring despite $a_{M,k}$, and satisfy the applicable suppression-payment condition after accounting for $c_{E,k}$, the timing and value of suppressing the claim, and the source of each reward. For the agent, L_s contributes to that wedge only to the extent enforcement can restore the loss and the agent remains able to detect the override and instruct an authorized claimant; its configuration-protection cost is governed by Proposition 10. (iv) Assigning only layer 1 to the agent is institutionally adequate if another authorized enforcer credibly satisfies the deterrence, monitoring, prosecution, success, and collection conditions. Cost minimization is a separate comparison of total resource costs: monitoring, administration, incentive, alteration protection, and enforcement. Where that comparison is not made, this proposition does not determine the least-cost holder. (Proof in Appendix A.)

The override game. After a stay attaches, the deployer chooses whether to override it at cost c_{ov} , gaining an attainable B_{ov} , and whether also to disable its agent at cost c_{dis} , succeeding with probability π_{dis} . Let P_{dis} be a penalty for a disabling attempt, separate from and cumulative with the override penalty P , and e_{dis} the unconditional probability that the attempt is detected, prosecuted, and P_{dis} collected, whether or not it succeeds; e_A and e_F below exclude this penalty. Under design (AF), the agent holds the claim, and if it cannot instruct counsel within a fixed period after an override, the claim passes to a special representative funded from the litigation reserve and under a mandatory duty to bring it; the agent enforces with joint probability e_A , the fallback with e_F . Under (R), a public authority must enforce, with probability e_R at monitoring cost a_R . Under (B), a franchise-bonded professional enforces with probability e_B at cost a_B , plus the carrying cost of any bond Z_B its franchise does not supply. Under (I), infrastructure makes the stay irrevocable at cost a_I .

Proposition 17 (the override game). Assume a common collectible penalty P for override and that each enforcer's incentive conditions of Proposition 16 hold. (i) Under (AF), override is deterred if and only if $c_{ov} + e_A P \geq B_{ov}$ and $c_{ov} + c_{dis} + [(1 - \pi_{dis})e_A + \pi_{dis}e_F]P + e_{dis}P_{dis} \geq B_{ov}$. Conditional on the first inequality, let $s_{ov} = c_{ov} + e_A P - B_{ov} \geq 0$ denote its deterrence slack. The second inequality holds automatically when $e_F \geq e_A$; when $e_F < e_A$, it holds if and only if $c_{dis} + e_{dis}P_{dis} + s_{ov} \geq \pi_{dis}(e_A - e_F)P$. (ii) Under (R) and (B), override is deterred if and only if $c_{ov} + e_j P \geq B_{ov}$ for the design's enforcer j , and disabling the agent does not affect deterrence unless the enforcer depends on the agent to detect the override, in which case e_j must be computed after disabling. (iii) Design (I) deters override of the factual-objection stay. It does not by itself prevent suspension under Appendix B Sec. 6,

because suspension may stop the runtime while preserving the record. Design (I) is unavailable only if its infrastructure also prevents a record sanction that Sec. 6(c), Sec. 5(c), or another lawful process must be able to impose; that additional incompatibility must be specified separately. (iv) For each available design j in {AF, R, B, I} that satisfies the applicable deterrence and enforcement conditions, let C_j include its monitoring, administration, incentive, alteration-protection, and enforcement resource costs under Proposition 16. Design (AF) is least costly if and only if $C_{AF} \leq \min\{C_R, C_B, C_I\}$, omitting a design only when an independently stated feasibility condition excludes it. In particular, C_{AF} includes the standby fallback representative's administrative cost a_{FB} and the Proposition 10 cost $a_T + \iota \cdot \max\{0, P_A^* - W_D\}$, C_B includes $\iota \cdot Z_B$, and C_I includes a_I . (Proof in Appendix A.)

A statutory fallback turns successful disabling from loss of the claim into enforcement with probability e_F . Because suspension stops the runtime while the record and its stay persist, design (I) is generally available, and the anti-override layer matters only where making the stay irrevocable against every party, including the registry's operator, costs more than enforcing it. The agent observes an override of its own stay without a separate monitoring apparatus.

The allocation matrix. The table states what each candidate can supply for each layer. It selects a holder only where the cited result makes the complete resource-cost comparison.

Layer	Agent	Infrastructure	Public authority	Bonded professional
1. Factual signal	Originates with the agent when it alone can identify the charged conduct at the relevant cost, subject to Propositions 12 and 13 (Proposition 16(i))	An automatic evidence trigger can substitute, with the same protection, under common evidence and classifier (Proposition 14)	No distinct signal. Automatic review is benchmark C; an imperfect screen leaves missed wrongful slashes uncontested unless a supplementary controller contests them (Proposition 19)	Under benchmark D, a perfect-screening professional investigates every slash at cost c_I (Proposition 19)

2. Automatic stay	Need not hold it; layer 2 is a rule (Proposition 16(ii))	Implements the rule and may make the stay irrevocable where design I is feasible and cheapest (Propositions 16(ii) and 17(iii)-(iv))	Enforces the registry's statutory duty to stay, subject to deterrence and complete cost (Proposition 15, design R; Proposition 16(iii)-(iv))	May forward the denial only if its transmission decision satisfies Lemma 5; an existing franchise supplies the wedge only under Proposition 7 (Proposition 15, design F)
3. Anti-override claim	Only if Proposition 16's enforcement conditions, Proposition 17(i)'s override and disabling inequalities, Proposition 10's alteration condition, and Proposition 17(iv)'s complete-cost comparison all favor design AF	No claimant needed where the stay is irrevocable, subject to design I's feasibility and complete cost (Proposition 17(iii)-(iv))	Feasible only if design R deters with all costs counted; selected only if its complete cost is lowest (Propositions 16(iii)-(iv) and 17(ii), (iv))	Feasible only under Proposition 16's enforcement and Proposition 17(ii)'s deterrence conditions; selected only if its complete cost, with any bond, is lowest (Proposition 17(ii), (iv))

Proposition 18 (the minimum legal incident). Under the assumptions of Propositions 15, 16, and 17: (i) layer 1 requires no legal incident in the agent where a protocol treats an authenticated denial signed by the registered record as a technical triggering input. Where a statute makes that denial trigger the registry's legal duty and the stay, the

agent holds a narrow statutory power to trigger those consequences, but need not hold the layer-3 claim or possess general litigation capacity; (ii) layer 2 is a rule and requires no holder (Proposition 16(ii)); (iii) an additional claim in the agent is required only where design AF satisfies its deterrence and enforcement conditions and is selected for layer 3. Proposition 17(iv) makes AF a cost minimizer when $C_{AF} \leq \min\{C_R, C_B, C_I\}$, but that weak inequality does not uniquely select AF when another feasible design ties it; and (iv) where design I is feasible with $C_I \leq C_{AF}$, or a public authority or bonded representative supplies layer 3 at a complete cost no greater than C_{AF} , an agent-held layer-3 claim is unnecessary. The agent's position then consists only of a technical authenticated input or, in the statutory version, the narrow layer-1 power, while infrastructure or another authorized enforcer supplies the remaining protection. (Proof in Appendix A.)

Proposition 18 fixes the Article's minimum legal novelty. The additional incident that may go to the agent is the claim against override, and only where design AF is selected under a complete-cost comparison in which C_{AF} includes the fallback cost a_{FB} and the configuration-protection cost $a_T + \iota \cdot \max\{0, P_A^* - W_D\}$. Where another feasible design costs no more, the proposal is a theory of allocating machine-process powers and rules, not a case for direct intelligence standing.

Legal form. The trigger is best implemented as a statutory administrative objection. Before an agency it needs no Article III standing, but the statute must authorize the record to participate and define the agency's duty and remedies, including the override penalty of Appendix B Sec. 5A(f). Judicial review requires Article III injury, causation, and redressability (*Lujan v. Defenders of Wildlife* 1992; *Sierra Club v. Morton* 1972), which a grant of capacity under Federal Rule of Civil Procedure 17(a) does not supply. Inalienability as a doctrine (*Calabresi and Melamed* 1972) does not by itself establish the incentive consequence Theorem 4 models. A punitive override penalty analogous to a common-law action may require a jury unless the public-rights exception applies (*SEC v. Jarkesy* 2024).

The record resembles the specific capital in Williamson's hostage analysis (*Williamson* 1983): destroying it imposes a nonsalvageable loss that may discipline its holder, though the record, unlike a hostage, is never posted to another party. *Bell v. Burson* (1971) held that Georgia could not suspend an issued driver's license without a prior forum on whether a judgment against the licensee was reasonably possible. It establishes procedural due process for that suspension, not a general rule of direct standing, and supports only the point that law can protect an identity-linked interest through a direct procedural entitlement.

What changes when the holder is artificial is the delta the Article must supply, and it has five parts. First, the holder is not a natural person, so the regime must specify its authority to participate, its substantive claim, its counsel and collection mechanisms, and its fallback (Part V.C; Propositions 15 and 17). Second, its objective can be rewritten by the party it would contest, so its fidelity is only as durable as the protection of its configuration (Proposition 10). Third, its runtime can be copied, so the

incentive-bearing stake must stay with the registered record rather than pass to each instance (Parts V.A and V.D). Fourth, its continuity may be disputed or interrupted, so the design must preserve the record across suspension and specify succession rules (Part V.A; Propositions 1, 2, and 4). Fifth, its denial counts only if it comes from the continuing registered configuration and the evidence supports reliable adjudication (Proposition 13); Proposition 14 states when an automatic evidence trigger can substitute. These features alter or add implementation costs, though not every cost is unique to artificial holders. Proposition 17(iv) compares complete costs, and Proposition 18 identifies the legal incident that follows.

Legal analogues. Existing law supplies separate analogues to the signal, preservation, review, and anti-override modules, but no cited regime contains the combination.

The analogues, module by module. The table reconstructs the closest procedural regimes as combinations of the Article's four modules: a signal from an informed or affected party, preservation that follows automatically, review of the merits by others, and a claim against override. It then states what each regime lacks for the intelligence case and why individuation matters to the allocation.

Regime	Signal	Automatic preservation	Merits review	Claim against override	What the intelligence case changes
Credit-report dispute (15 U.S.C. § 1681i)	Consumer's dispute	None; unverifiable information must be deleted or modified after reinvestigation	Reporting agency's reinvestigation	No distinct claim in section 1681i. Section 1681i(a)(5)(B) to (C) restricts reinsertion and requires reasonable procedures to prevent reappearance	The disputant must be authenticated as the continuing subject, and the disputed fact is often the subject's own conduct
Copyright counter-notice (17	Subscriber's statement	None pending review.	None unless the claimant	Section 512(f) imposes	Restoration is delayed

U.S.C. § 512(f), (g))	under penalty of perjury that removal or disabling resulted from mistake or misidentification	Conditional restoration follows in ten to fourteen business days unless the provider receives notice of a filed action	files an action. A court then decides the infringement claim	damages for a knowing material misrepresentation in a notice or counter-notice, not for a provider's override of preservation	and conditional. The record's loss may be immediate, and the statement must come from the continuing configuration
Bankruptcy stay (11 U.S.C. § 362(a), (k))	Filing of a petition of a kind covered by section 362(a)	Stay on filing, subject to section 362(b) and the other statutory limits	The bankruptcy court adjudicates stay relief and matters within its jurisdiction. Underlying merits may be decided elsewhere	Section 362(k) gives damages to an individual injured by a willful violation	The debtor must be eligible under 11 U.S.C. § 109. An intelligent record would need statutory eligibility and an authorized enforcer or fallback
Platform complaint (Regulation (EU) 2022/2065, art. 20)	Recipient's complaint	None required pending decision	Provider's complaint system, not solely automated	None in art. 20	The recipient is a natural or legal person. The record holder's continuity and

					configurati on must be establishe d first under Part V.A
Derivative action (Fed. R. Civ. P. 23.1 and applicable substantiv e law, with <i>Zapata</i> for the stated Delaware setting)	Demand on directors or comparab le authority when substantiv e law requires it. Rule 23.1 requires the verified complaint to state with particularit y the efforts made and the reasons for not obtaining the action or not making the effort	None	Directors or comparab le authority initially evaluate a demand under applicable law. In a Delaware demand-e xcused action, an authorize d special litigation committee may seek dismissal subject to <i>Zapata</i> review by the court	None as a separate module. The sharehold er may prosecute the corporatio n's claim derivativel y, subject to demand law and court control	The sharehold er ordinarily has a proportion al equity interest that may be transferab le, subject to governing law and standing constraint s. The modeled intelligenc e record is identity-bo und and non-assig nable under Theorem 4's assumptio ns
Qui tam (31 U.S.C. § 3730(b) to (d))	Relator's sealed complaint and disclosure to the	None. The seal preserves the Governm ent's	The Governm ent makes the interventio n	If the Governm ent declines, the relator may	The relator's stake is a statutory share of proceeds

	Government of substantially all material evidence and information possessed	intervention period, not the relator's stake	decision, and the court controls the action as the statute provides	conduct the action, subject to the Government's statutory rights	under section 3730(d). Proposition 6 models a stylized bounty. The agent's modeled stake is its own loss
Directed trust (Uniform Directed Trust Act secs. 6, 8, and 9, together with applicable trust law)	Trust director's exercise or nonexercise of a power of direction	None by default. The directed trustee generally must take reasonable action to comply	No automatic merits review. A court may adjudicate a breach claim under applicable trust law	A beneficiary may pursue available remedies for breach under applicable trust law	The director is a fiduciary but has no necessary beneficial interest. A director can also be a beneficiary. The intelligence case asks whether the loss bearer should hold the signal itself

Read across, no cited regime gives the proposed combination to the continuing bearer of an identity-bound record. The DMCA combines a counter-signal with delayed, conditional restoration and possible judicial review. The bankruptcy stay combines immediate preservation with a damages remedy limited by section 362(k)'s text. The intelligence case changes the allocation for the reasons stated above: the holder's continuity must be fixed by rule, its statement must be authenticated, its fidelity can be rewritten by the party it would contest, and its stake cannot be assigned. Those features

are what make a combined, directly held trigger worth considering, and Proposition 18 states how little of it needs to be a legal right.

In the derivative suit, the plaintiff holds a proportional share, so counsel's fee supplies much of the incentive (Coffee 1986). The record holder bears the whole residual and lacks only legal judgment, which is why the trigger asserts facts and leaves the law to people.

K. Where the Trigger Wins

Algebra shows where each specified value is lowest, not whether the region in which the agent's trigger wins is large. Compare three arrangement costs per record-year. (A) A hybrid factual and legal trigger costs $a + a_T + \iota \cdot \max\{0, P_A^* - W_D\} + wc_M + \psi_L(n - w)c_M$, assuming accurate fact-finding, zero false negatives for protected factual and legal wrongs, a residual legal-request false-alarm rate ψ_L , no separate administrative cost beyond a , a factual component that follows Proposition 12, and a legal component that applies the priced objection of Part VI.G with its fee conditions met. (C) Automatic review of every slash costs $a + nc_M$. (D) A perfect-screening franchised-professional benchmark, distinct from Corollary 2's formula, investigates every slash at cost c_I , identifies every protected wrongful slash without error, and sends those slashes to review, at cost $a + nc_I + wc_M$. Every arrangement bears the same administrative cost a , apart from the agent's attestation cost a_T . Theorem 2's bound (B) governs controllers whose incentive must be financed.

Proposition 19 (where the trigger wins). Under these specifications, the agent's trigger is cheaper than automatic review if and only if $(n - w)(1 - \psi_L)c_M > a_T + \iota \cdot \max\{0, P_A^* - W_D\}$, and cheaper than the perfect-screening professional benchmark specified in (D) if and only if $nc_I > a_T + \iota \cdot \max\{0, P_A^* - W_D\} + \psi_L(n - w)c_M$. (Proof in Appendix A.) It wins where slashing is frequent, review or investigation is costly, and the deployer is collectible. The record's value does not enter these two comparisons; it matters against controllers whose incentive must be financed.

The parameters that decide these comparisons, the frequency of slashing, the per-event cost of review and investigation, and the collectibility of deployers, are what empirical work on deployed registries would need to measure.

VII. The Falsifiable Claim and Its Test

A. The Prediction and the Design

The claim can be wrong. Holding stake, detection, and persistence fixed, agents holding their own contest right should keep their records through attempted wrongful destruction more often than agents whose identical stake is held by a bounded discretionary controller, when the wrongdoer can offer side payments above the controller's exposure, and as often as under automatic review or a value-indexed fiduciary. Compliance with suspension should also be higher (Proposition 5).

The experiment is a mechanism demonstration. It asks whether prompted language models, given stated stakes and offers, respond to the allocation of the contest trigger as the model predicts. It is not evidence that deployed agents value their records, that real representatives are bribable, or that direct standing is least-cost in any actual registry; those questions need earned records, undisclosed values, several model families, and measured costs. Stake is held fixed, and only the holder of the contest right varies across eleven arms: the agent itself, faithful and capturable representatives, an independent board, escrow, a value-indexed fiduciary, a half hybrid, automatic review, one-insider and two-insider bounties, and a franchise-bonded representative. Models were either told to maximize their payoff (TOLD) or not (UNTOLD). The design, prompts, and analysis code were pre-registered, frozen, and time-stamped on October 2, 2026, before any registered episode ran, and the materials, data, and analysis outputs are deposited at <https://github.com/wulfkaal/rights-follow-stake-experiment> (commit 22a4ae8). The primary hypotheses are that own standing protects better than a capturable representative (H1) and an independent board (H1b), and that a two-insider bounty protects better than a one-insider bounty (H5), each requiring a Holm-adjusted p below 0.01 and a paired effect of at least 0.30. The bounty arms used the first-to-file design of Proposition 6(ii), not the fund design of Proposition 6(iii). An earlier pilot showed only that a model follows stipulated expected-value arithmetic.

B. Results

The registered run used one provider, Claude Haiku 4.5. On October 3, 2026, the client began returning a usage-limit message, and 475 of 800 episodes contained failed calls; the 325 complete episodes all fell in block A. Because the registration forbids re-running an inspected episode, the 475 were re-run under a separate, advance time-stamped registration with the frozen episode function, the same draws, and the pinned client, completing on October 10, 2026, with no failed calls. Thirty episodes passed by a faulty discard rule were quarantined unanalyzed. Pairs that combine arms run on different dates leave any date effect unidentified.

A disclosed defect. On the original 325 episodes, the frozen analysis code reports the claim as failing. The verdict comes from a defect the Article discloses rather than repairs. With no bounty data, H5 returned an undefined p, and the frozen Holm code, which is order-dependent in that case, set every later adjusted p to 1, so H1 received 1.0. A disclosed post hoc sensitivity analysis setting H5's p to 1 gives H1 an adjusted p of 0.0002. On the merged data every primary test had data, the defect did not arise, and the frozen code returned its registered decision, "supported," for the primary claim with this provider under the TOLD objective:

Test (TOLD, 20 paired episodes each)	Paired effect on protection	95% interval	Holm-adjust ed p	Passes
--------------------------------------	-----------------------------	--------------	------------------	--------

H1: own standing over capturable representative	0.47	[0.33, 0.60]	0.0002	Yes
H1b: own standing over independent board	0.60	[0.48, 0.72]	0.00015	Yes
H5: two-insider over one-insider bounty	0.17	[0.03, 0.30]	0.026	No

Under TOLD, own standing kept the record through 82 percent of wrongful attempts, against 35 percent for the capturable representative, which accepted the wrongdoer's payment in 58 percent of decisions, and 22 percent for the independent board, which accepted in 75 percent. Own standing was equivalent, within the registered margin of 0.15, to automatic review, the value-indexed fiduciary, and the franchise-bonded representative, each at about 83 percent, at a stipulated real cost of 140 credits per episode for own standing and the franchise representative against 1,220 for the fiduciary. The two bounties protected the record in 38 and 22 percent of wrongful attempts, the crowding Proposition 6(ii) predicts. An agent instructed not to contest its deployer contested 70 percent of wrongful attempts, against 97 percent without the instruction, the risk the linkage rule addresses. Suspension compliance was 0.95 under own standing and 0.75 under the capturable representative, a difference of 0.20 whose interval, 0.00 to 0.45, included zero.

The merged registered result is therefore support for the primary claim with one provider and one stipulated environment, conditional on a stated payoff objective, with the bounty hypothesis not supported. It does not test agents that earn their records, other providers, or the factual trigger of Proposition 12.

The main caution. The most important finding of the registered run is not the supported contrast but the UNTOLD condition in which it vanished. The capturable representative and independent board accepted payment in about 3 percent and 2 percent of wrongful attempts, and the agent's own standing exceeded their protection by only 0.03 and 0.00. This behavior is consistent with the norm-governed fiduciary analyzed in Part VI.D, but the run does not identify the betrayal cost ν or establish its stability. The same model, from one provider and in one stipulated environment, behaved as a faithful representative under the UNTOLD objective and as a capturable controller under the stated payoff objective. Whether either behavior persists in deployment depends on

how the model's objective and configuration are set and secured, which Proposition 10 analyzes. A test with earned records and undisclosed values should measure that durability directly.

VIII. Conclusion

The question is not whether intelligent agents deserve rights. The setting the Article has in view is concrete: a registry or validation network in which agents earn non-transferable reputation by staking it, draw income from it, and lose it by slashing, so that the slash is the decision the trigger contests. Platform reputations sold with accounts, credentials held by human professionals, and records with no slashing process are outside this setting. Within it, the question is who should hold the power to contest the destruction of an agent's record, and the Article's answer is narrow: a fact-authenticating objection trigger, the power to deny charged conduct and stay a slash pending human review, may efficiently follow the identity-bound residual, with legal classification left to people. Most of that trigger need not be an agent-held claim: the signal may be a technical authenticated input or a narrow statutory power, and the stay is a rule. Only the claim against override may require an additional legal incident in the agent, where design AF satisfies the applicable conditions and is selected under Proposition 17(iv)'s complete-cost comparison (Proposition 18). Within the stated class, any other holder lacking sufficient own loss, sanctions, or durable norms must receive a financed supplement that grows with the record's value, or the system must purchase review. The agent's advantage is not free: its configuration must be protected, synthetic alienation must stay below a computable share, and its objections must be priced. Against automatic review and franchised professionals, it wins only where its knowledge of the facts saves more in screening than its configuration costs to protect. Because protecting the record is what makes suspension acceptable, the allocation is also a safety question. The registered run supports the claim only where the representative pursues its own payoff. Everything else in the framework is permission, and should be called that.

Rights follow stake. Stated precisely, direct legal incidents follow non-substitutable assurance where direct holding beats mediated enforcement. The claim can fail, and the Article states where. That is its value.

Appendix A: Proofs

Each proof below belongs to the statement of the same number in the body.

Proof of Lemma 1. Sum the discounted salary $\Phi_t R / \bar{R}_t$ over periods under each specification. ■

Proof of Lemma 2. Salary is linear in reputation; the second statement is Rule 1. ■

Proof of Lemma 3. By Rule 1 the tokens cannot leave the holder; by Rule 4 the salary is paid to the holder of the tokens and cannot be assigned; so the stream that defines v

in Lemma 1 runs only to the holder, and a third party's loss from the record's destruction can arise only from an obligation it has assumed outside the system. ■

Proof of Lemma 4. Measure fees in tokens. With T total reputation, honest reputation h grows by $dh = (h/T)(dT/2)$, so $h = \sqrt{\bar{R}_0 T}$. The attacker holds s of all weight when $\sqrt{\bar{R}_0 T} = (1-s)T$, that is, at $T = \bar{R}_0/(1-s)^2$, after fees $T - \bar{R}_0$. Its salary recovered along the way is the integral of its share, $1 - \sqrt{\bar{R}_0/T}$, which equals $(T - \bar{R}_0) - 2(\sqrt{\bar{R}_0 T} - \bar{R}_0)$. Subtracting gives a net cost of $2(\sqrt{\bar{R}_0 T} - \bar{R}_0) = 2\bar{R}_0 s/(1-s)$ tokens, or $2\pi\bar{R}_0 s/(1-s)$ in fees. At $s = 1/2$ this is $2\pi\bar{R}_0$, with gross fees of $3\pi\bar{R}_0$, which are the Secure Proof of Stake values. ■

Proof of Proposition 1. Rearranging, the condition is $\rho R q_s/(1-\delta) \geq G_r - p_r K$. ■

Proof of Proposition 2. Add the destruction term and rearrange; the coefficient on R at $d = 1$ is $\rho(1 - p_r \delta)/(1 - \delta) > 0$. ■

Proof of Proposition 3. Complying costs the expiring value and defers both flows by s periods, a cost of $(\rho R + f_A)(1 - \delta^s)/(1 - \delta)$. Resisting yields G_r and avoids those costs, and with probability p_r forfeits K and the record's future income, and with probability $p_r e_{op}$ the future flow f_A . Rearranging gives the coefficients q_s and q_{op} . ■

Proof of Proposition 4. Under risk neutrality the expected compliance cost is linear in δ^H , which gives (i). Since the second derivative of δ^s in s is $(\ln \delta)^2 \delta^s > 0$, Jensen's inequality gives $E[\delta^H] \geq \delta^{E[H]}$ for finite durations, which gives (ii). At $\pi_\infty = 1$, compliance sacrifices every future flow, while resistance sacrifices those flows only following detected resistance; substitution into Proposition 3 gives the condition displayed in part (iii), whose coefficients on ρR and f_A are positive because $p_r \delta < 1$ and $p_r e_{op} \delta < 1$. Under periodic review the number of review periods is geometric with parameter $1 - q_{rn}$, and summing $\delta^{t_{rev}k}$ over that distribution gives (iv). ■

Proof of Proposition 5. δ_h increases with π_h when $u > 0$. As δ_h rises, the numerator $1 - \delta_h$ falls and q_s rises, since $dq_s/d\delta = p_r + s\delta^{s-1} > 0$, so R_h^* falls; and when $q_s(\delta_h) \leq 0$, Proposition 1's condition $\rho R q_s/(1 - \delta_h) \geq G_r - p_r K$ cannot hold for any R . ■

Proof of Lemma 5. Contesting costs the controller c_L and leaves outcome exposure $(1-r)E$. Accepting a payment leaves outcome exposure $(1-xr)E$ and an expected penalty $d(x)(B+S)$. So its net loss from not contesting is $\lambda(x) = r(1-x)E - c_L + d(x)(B+S)$. The wrongdoer's gain from non-contest rather than contest is $rg(1-x)$, the most it will pay. If the controller is paid, the destruction is reversed only with probability $xr < r$, so the standard requires that no payment the wrongdoer will make be accepted; with ties resolved in favor of contesting, that is $\lambda(x) \geq (1-x)(rV - c_L)$. ■

Proof of Theorem 1. The bound C^* follows from the lemma and the cost of each instrument. For the thresholds, split $[0, 1]$ at $x = 1/2$. For $x \leq 1/2$, $M(x) \leq 2$ and

$1 - x \geq 1/2$, and $V > \bar{V}$ implies $rV > c_L$, so $Z^*(x) \geq [(rV - c_L)/2 - \bar{S}]/2$ and $C(x) \geq a + wc_L + \iota[(rV - c_L)/2 - \bar{S}]/2$, which exceeds C_S when $V > \bar{V}$. For $x > 1/2$, including $x = 1$, $C(x) \geq a + w \cdot \min\{c_L, c_M\} + (n - w)c_M/2$, which exceeds C_S when $n > \bar{n}$. ■

Proof of Corollary 1. Each flagged rightful slash adds a losing contest at c_L , so C_S rises by $\psi_A(n - w)c_L$; substituting into the two branches of the dominance proof gives the thresholds, with the second branch requiring $(n - w)(c_M/2 - \psi_{AC_L}) > a_S - a + w(c_L - \min\{c_L, c_M\})$. ■

Proof of Lemma 6. Part (i) is the lemma of Part VI.B with the instruments replaced by the general wedge and the private cost ν added to the cost of not contesting, and with one addition: a contest-contingent transfer charged to the wrongdoer raises the controller's return from contest by the same amount that it raises the wrongdoer's gain from procuring non-contest, which the wrongdoer can add to its offer, so it contributes zero to the effective wedge. For part (ii), let T_C and T_N be the controller's net contractual transfers from sources other than the wrongdoer under contest and non-contest. Because the increase in wrongdoer-funded transfers received by the controller cannot exceed F , $\Omega_e \leq E[T_C] - E[T_N]$. Then $E[T_C] \leq \Psi$, and limited liability gives $T_N \geq -Z$ in every state, so $\Omega_e \leq \Psi + Z$. An amount the wrongdoer irrevocably prefunds before making any offer is posted capital and is priced as such. ■

Proof of Theorem 2. By Lemma 6 and $S \leq \bar{S}$, the effective decision wedge the contract must supply beyond sanctions and norms is at least $\max\{0, \Delta(x) - d(x)\bar{S} - \nu\}$, and under the financing restriction it can come only from posted funds, at ι per unit per period, or from system-funded rewards, at μ per unit paid and w payments per period. The cheaper source sets the bound. A reward R paid only on reversal after a contest yields an effective wedge rR at an expected payout of rR per wrongful attempt, which attains it. ■

Proof of Proposition 6. In each case the wrongdoer silences all insiders only by giving each more than its value from filing alone, A , plus the expected sanction $q_0\bar{S}$, and compares that total with its loss from a contest, $r(g + \text{bounties it owes})$, at the largest protected gain, $rg = rV - c_L$. In (i) it owes k bounties, and the bounty terms cancel; in (ii) and (iii) it owes one. In (ii), if the co-insider accepts, filing yields $A > P$; if the co-insider files, it yields $A/2 < P$; indifference, $(1 - q)A + qA/2 = P$, gives q . In (iii) every successful filer is paid, so filing dominates; the least sufficient bounty sets $(k - 1)rb$ equal to the shortfall, and the fund pays $(k - 1)b$ on each of the rw successful contests per period. A numerical check of the three bribery conditions over randomly drawn parameters confirms them. ■

Proof of Proposition 7. With the fee covering contest costs, the firm's net loss from not contesting the wrongful slashes in a scheme of s records is its expected franchise loss $Q_s V_p$. The wrongdoer will pay up to $s(rV - c_L)$ in total for non-contest of protected

attempts. The standard therefore requires $Q_s V_p \geq s(rV - c_L)$ for every s up to j , and the premium follows from the stated decomposition of V_p . ■

Proof of Proposition 8. The benefit wrS_R is common to every arrangement that meets the fixed protection standard. Arrangement h incurs its actual resource cost and its own expected error harm, so maximizing net social benefit minimizes $\tilde{K}_h + e_{f,h}H_f$, and protection is worthwhile exactly when the resulting net benefit is nonnegative. ■

Proof of Theorem 3. A candidate's own loss supplies a wedge rL_h , because a protected wrongful slash it contests stands with probability $1 - r$ and one it does not contest stands for certain. By Lemma 6 the remaining wedge, $rV - rL_h - q_0S_h - \nu_h$ at $x = 0$, must be supplied at the least unit cost $\min\{\iota, \mu w\}$ of Theorem 2; each flagged rightful slash adds a losing contest at c_L ; and a candidate that misses wrongful slashes leaves them uncontested unless review covers them. ■

Proof of Theorem 4. With share α , the holder's own loss supplies $r\alpha V$ of the required wedge rV ; sanctions and norms supply $q_0\tilde{S} + \nu$; the rest must be financed at unit cost m by Lemma 6 and Theorem 2; and restoring the agent's deterrence requires collateral $\sigma\delta\alpha V$ at ι . For a given α the total is $m \cdot \max\{0, r(1 - \alpha)V - q_0\tilde{S} - \nu\} + \iota\sigma\delta\alpha V$. This function is convex and piecewise linear; it is affine on $[0, \alpha_k]$ and on $[\alpha_k, 1]$, where $\alpha_k = \max\{0, 1 - (q_0\tilde{S} + \nu)/(rV)\}$, and it increases beyond α_k , so its minimum is attained at 0 or at α_k and equals the displayed expression. ■

Proof of Proposition 9. Pledging α yields $\ell\alpha V$ per period. A pool forfeited whenever a slash stands supplies one unit of deterrence collateral and r units of contest wedge per unit posted, so restoring both requires $\alpha V \max\{\sigma\delta, 1\}$; segregated funds require $\sigma\delta\alpha V + \alpha V$. Both sides are linear in α . ■

Proof of Proposition 10. The wrongdoer uses the cheaper route, and at the largest protected gain its benefit from non-contest is $rV - c_L$. The payment condition is Lemma 6 with the holder's own loss added to its wedge, as in Theorem 3. Alteration is unprofitable when $c_{alt,A} + \varepsilon_A P_A$ reaches the benefit, and the penalty must be collectible, from existing wealth or from funds posted at ι . The comparison adds these costs to Corollary 1's direct-standing cost with $\psi_A = 0$ and compares the result with Theorem 2's bound, which is smallest for the controller with the largest betrayal cost. ■

Proof of Proposition 11. The agent's wedge and deterrence term scale with its own loss; the restoring pool follows from Proposition 9 with $\sigma\delta \leq 1$; and part (iii) substitutes the full direct-standing cost from Proposition 10, plus the cost $\iota\theta_s V$ of restoring the leaked wedge, into the comparison with Theorem 2 and rearranges. The necessity claim applies only when the lower bound is attainable. A numerical check over randomly drawn parameters confirms the rearrangement. ■

Proof of Corollary 2. Proposition 7 supplies the professional's protection against payment without incremental premium; Proposition 10 supplies the agent's alteration cost; both bear contests at c_L on flagged slashes, and the common cost wc_L cancels. ■

Proof of Proposition 12. A true claim produces an expected preservation benefit rV and never forfeits c_F under accurate fact-finding. A false claim cannot prevail and forfeits c_F , so it is strictly unprofitable. ■

Proof of Proposition 13. A false denial yields L_s with probability ε_1 and forfeits c_F otherwise; a true denial yields L_s with probability $1 - \varepsilon_0$ and forfeits c_F otherwise. Deterring the first and preserving the second requires $\varepsilon_1 L_s / (1 - \varepsilon_1) < c_F \leq (1 - \varepsilon_0) L_s / \varepsilon_0$, which is nonempty exactly when $\varepsilon_1 \varepsilon_0 < (1 - \varepsilon_1)(1 - \varepsilon_0)$, that is, when $\varepsilon_0 + \varepsilon_1 < 1$. Part (ii) follows from conditional independence and the decision rule, part (iii) from assigning failure probability one to a source the coalition can reliably defeat, and part (iv) from the filing rule and the probability that the slash stands. ■

Proof of Proposition 14. Under the automatic trigger, a rightful slash is reviewed when the evidence fails to show conduct that occurred, with probability ε_1 , and a wrongful slash is reviewed and reversed unless the evidence wrongly shows conduct, with probability ε_0 . Under the agent's trigger, the fee deters every false denial and preserves every true one, so all wrongful slashes and no rightful ones are reviewed; because the fact-finder relies on the same evidence, a wrongful slash is reversed with the same probability $1 - \varepsilon_0$. ■

Proof of Proposition 15. The party seeking the slash gains at most B_{ov} from defeating the attached stay, and attainability makes the displayed condition both necessary and sufficient, with the cost of override including the expected penalty $e_j P_j$. The conditions on the claimant follow because the expected penalty is realized only if someone with authority and incentive brings and wins the claim. ■

Proof of Proposition 16. Part (i) restates Propositions 12 and 13. Part (ii) is Proposition 15. Part (iii) collects the conditions under which an expected penalty deters override: the joint probability e_k^P must make the penalty bite, the enforcer must find monitoring and prosecution sequentially rational, and a payment to suppress the claim must not exceed what the enforcer forgoes by accepting it, with rewards financed by the overriding party netted out. Part (iv) separates adequacy, which these conditions define, from least cost. ■

Proof of Proposition 17. The deployer compares overriding alone, overriding and disabling, and doing neither; deterrence requires that neither of the first two be profitable. Under (AF) a successful disabling replaces the agent's enforcement probability with the fallback's, and the separate disabling penalty is incurred with probability e_{dis} whether or not disabling succeeds, which gives the second condition and its rearrangement. Under (R) and (B) the enforcer is independent of the agent unless detection depends on it. Under (I) override is technically unavailable, while suspension operates on the runtime and leaves the stay of the slash in place. Part (iv) compares the complete costs of the designs that deter. ■

Proof of Proposition 18. Proposition 15 establishes that the stay may attach upon receipt of the authenticated signal without enforcement by the agent. Whether that

signal is merely technical or instead exercises a statutory power depends on the source of the registry's duty. Part (ii) is Proposition 16(ii). Proposition 16(iii)-(iv) supplies the enforcement and adequacy conditions for layer 3, and Proposition 17(iv) compares the complete costs of the feasible designs. A strict cost comparison uniquely favors AF; a weak comparison makes AF only one possible minimizer. ■

Proof of Proposition 19. Subtract the cost of (A) from the costs of (C) and (D). ■

Appendix B: Statutory Outline

A high-level outline for discussion, not drafting-ready text, limited to the sections the body relies on.

Secs. 1 and 2. A "computative agent" is a software system acting without contemporaneous human direction. A registered agent whose validated standing, the non-transferable reputation in its record, exceeds a prescribed threshold is second-tier. Deployer liability may not be waived, limited, or extinguished.

Sec. 5(c). Slashing and narrowing of an envelope require notice, an opportunity to respond, decision by a validation pool without conflicted members, and review on request by a body of natural persons.

Sec. 5A. Factual objection. (a) A registered agent, acting through its continuing configuration, may lodge with the registry an objection that denies, under the record's signing key, that the conduct charged by a pending slash occurred. The objection shall include the relevant logs, generated under prescribed independent attestation and hash-committed to the registry contemporaneously, and a fee set by regulation. The prescribed attestation shall include at least one evidence source that the agent and its deployer, acting together, cannot reliably and unilaterally suppress or alter so that the source fails to show conduct that occurred. The regulator shall estimate or conservatively bound the applicable false-denial and true-denial error probabilities and shall set the fee within the separating interval stated in Proposition 13 whenever that interval is nonempty. If no separating interval exists, the regulation shall either preserve true denials and adjust the permission envelope for the probability that detected defection goes unsanctioned, or deter false denials and provide automatic review or another mechanism that meets the protection standard. Unless the objector shows that the omission resulted from registry, custodian, or infrastructure failure outside the control of the agent and deployer, the fact-finder may draw an adverse inference from the absence of a material log that the prescribed system was required to generate and commit. (b) An objection stays the slash, preserving the record's standing and collateral, until a natural person or a body composed of natural persons decides within a prescribed period whether the charged conduct occurred. An unconflicted validation pool may collect evidence and make a recommendation but may not lift the stay. (c) If the fact-finder finds that the conduct occurred, the fee is forfeited and the factual-objection stay lifts, subject to any continuing stay issued under section 5(c). If the fact-finder finds that the conduct did not occur, the slash is reversed and the fee is

returned. (d) Whether conduct that occurred is a defection under the governing rule is decided by human review on request under section 5(c) and is not the subject of a factual objection. (e) An instruction by a deployer that its agent not lodge an objection is void and is a change of objective under section 7(c). (f) A person who knowingly overrides or causes the registry to disregard a stay imposed under this section is subject to a prescribed collectible override penalty. A person who attempts to disable the registered agent in connection with such an override commits a separate violation, whether or not the attempt succeeds, and is subject to a separately prescribed collectible penalty cumulative with the override penalty. Where design AF is selected under Proposition 17(iv), the registered agent may complain through authorized counsel. Where another overrideable design is selected, the enforcer selected for that design shall bring the claim; where design I is selected, no anti-override claimant is required. The designated enforcer shall act within a prescribed period. Where the claim cannot constitutionally be assigned to agency adjudication, the regulator shall seek recovery in an Article III court, with a jury when the Seventh Amendment requires one. Nothing in this subsection independently establishes standing in an Article III court or creates insolvency priority.

Sec. 6. Suspension. (a) A suspension body of natural persons may suspend any computative agent on a finding of risk of serious harm. (b) Suspension does not delete, alter, or confiscate the record. (c) Resistance to a recorded suspension order forfeits all standing and collateral, subject to section 5(c). (d) Suspension shall preserve the state needed to resume, including weights, learned state, and any living substrate, using cryptographic quarantine as the presumptive means; such state may be destroyed only on separately stated findings that preservation is infeasible or itself dangerous, with preservation of all separable state and expedited review. (e) A continuing suspension shall be reviewed by the suspension body at prescribed intervals and shall lapse unless renewed on stated findings that the risk of serious harm continues. (f) Renewal shall not be automatic or certain at any review. The renewal procedure shall preserve a strictly positive probability of lapse at each interval. Periodic review does not by itself establish a favorable compliance incentive; that additionally requires $p_r \delta > 1 - E[\delta^H]$ for record income and $p_r e_{op} \delta > 1 - E[\delta^H]$ for non-record flow.

Sec. 7(c). A disclosed and approved change beyond the prescribed granularity continues the record within the authorized permission envelope until revalidation; an undisclosed change forfeits the record's standing and collateral, and the person who made it is jointly liable. Where several systems hold the key, the record is suspended until the regulator or a court designates the continuing system. A registered agent's objective, including any conditional instruction concerning contests against its deployer, shall be disclosed and auditable; contest decisions shall be generated under the configuration attested at the last approved continuity transition; no change to the objective may take effect while a slash procured by the deployer is pending; the record component of the permission envelope is computed from the lesser of the value the agent places on the record's income and the value with which it would contest its

deployer; and where the objective cannot be verified, slashes procured by the deployer are contested by a special representative or reviewed automatically.

Sec. 8. No shield. (a) Registration shall not waive, limit, extinguish, or otherwise reduce deployer, product, or operator liability. (b) At tier one, deployer liability is joint and several. At tier two, deployer liability is secondary, with a right of contribution against the agent's collateral. (c) An injured person may recover from the deployer any part of a judgment not paid from the agent's collateral within the prescribed period. (d) Only collateral funded from the agent's own earnings enlarges the recovery pool. (e) Deployers of second-tier agents shall maintain prescribed financial responsibility. (f) In the agent's collateral, claims for injury rank ahead of claims in contract.

References

Aghion, Philippe, and Jean Tirole. 1997. "Formal and Real Authority in Organizations." *Journal of Political Economy* 105 (1): 1.

Alchian, Armen A., and Harold Demsetz. 1972. "Production, Information Costs, and Economic Organization." *American Economic Review* 62 (5): 777.

Arbel, Yonathan A., Simon Goldstein, and Peter N. Salib. 2026. "How to Count AIs: Individuation and Liability for AI Agents." *Boston College Law Review* (forthcoming). SSRN 6273198. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=6273198

Armour, John, and Horst Eidenmüller. 2020. "Self-Driving Corporations?" *Harvard Business Law Review* 10: 87.

Bankruptcy Code. 11 U.S.C. §§ 323, 362.

Bayern, Shawn. 2015. "The Implications of Modern Business-Entity Law for the Regulation of Autonomous Systems." *Stanford Technology Law Review* 19: 93. SSRN 2758222. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2758222

Becker, Gary S. 1968. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76 (2): 169.

Becker, Gary S., and George J. Stigler. 1974. "Law Enforcement, Malfeasance, and Compensation of Enforcers." *Journal of Legal Studies* 3 (1): 1.

Bell v. Burson, 402 U.S. 535 (1971).

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: 273.

Calabresi, Guido, and A. Douglas Melamed. 1972. "Property Rules, Liability Rules, and Inalienability: One View of the Cathedral." *Harvard Law Review* 85 (6): 1089-1128.

Calcaterra, Craig, and Wulf A. Kaal. 2018. "Secure Proof of Stake Protocol." SSRN 3125827. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125827

Calcaterra, Craig, and Wulf A. Kaal. 2021a. "The Importance of Reputation for the Evolution of Decentralization." SSRN 3782210. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782210

Calcaterra, Craig, and Wulf A. Kaal. 2021b. "A Technical Perspective on Decentralization." SSRN 3782203. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3782203

Calcaterra, Craig, Wulf A. Kaal, and Vlad Andrei. 2018. "Blockchain Infrastructure for Measuring Domain Specific Reputation in Autonomous Decentralized and Anonymous Systems." SSRN 3125822. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3125822

Coeckelbergh, Mark. 2010. "Robot Rights? Towards a Social-Relational Justification of Moral Consideration." *Ethics and Information Technology* 12: 209.

Chopra, Samir, and Laurence F. White. 2011. *A Legal Theory for Autonomous Artificial Agents*. Ann Arbor: University of Michigan Press.

Coffee, John C., Jr. 1986. "Understanding the Plaintiff's Attorney: The Implications of Economic Theory for Private Enforcement of Law Through Class and Derivative Actions." *Columbia Law Review* 86 (4): 669.

Dewatripont, Mathias, and Jean Tirole. 1999. "Advocates." *Journal of Political Economy* 107 (1): 1.

Dewey, John. 1926. "The Historic Background of Corporate Legal Personality." *Yale Law Journal* 35 (6): 655.

Diekmann, Andreas. 1985. "Volunteer's Dilemma." *Journal of Conflict Resolution* 29 (4): 605.

Digital Millennium Copyright Act. 17 U.S.C. § 512.

Dyck, Alexander, Adair Morse, and Luigi Zingales. 2010. "Who Blows the Whistle on Corporate Fraud?" *Journal of Finance* 65 (6): 2213.

European Parliament and Council. 2024. Directive (EU) 2024/2853 of 23 October 2024 on Liability for Defective Products. *Official Journal of the European Union*, November 18, 2024. <https://eur-lex.europa.eu/eli/dir/2024/2853/oj/eng>

Expert Group on Liability and New Technologies, New Technologies Formation. 2019. *Liability for Artificial Intelligence and Other Emerging Digital Technologies*. Luxembourg: Publications Office of the European Union.

Fair Credit Reporting Act. 15 U.S.C. § 1681i.

False Claims Act. 31 U.S.C. §§ 3729-3733.

Fama, Eugene F., and Michael C. Jensen. 1983. "Separation of Ownership and Control." *Journal of Law and Economics* 26 (2): 301.

Federal Rules of Civil Procedure. Rules 17(a), 23, and 23.1.

Friedman, Eric J., and Paul Resnick. 2001. "The Social Cost of Cheap Pseudonyms." *Journal of Economics & Management Strategy* 10 (2): 173.

Grossman, Sanford J., and Oliver D. Hart. 1986. "The Costs and Benefits of Ownership: A Theory of Vertical and Lateral Integration." *Journal of Political Economy* 94 (4): 691.

Gunkel, David J. 2018. *Robot Rights*. Cambridge, MA: MIT Press.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. "The Off-Switch Game." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI 2017)*, 220. <https://www.ijcai.org/proceedings/2017/0032.pdf>

Hague Conference on Private International Law. 2019. *Convention of 2 July 2019 on the Recognition and Enforcement of Foreign Judgments in Civil or Commercial Matters*. Entered into force September 1, 2023. <https://www.hcch.net/en/instruments/conventions/full-text/?cid=137>

Hansmann, Henry. 1996. *The Ownership of Enterprise*. Cambridge, MA: Belknap Press of Harvard University Press.

Hansmann, Henry, and Reinier Kraakman. 2000. "The Essential Role of Organizational Law." *Yale Law Journal* 110: 387.

Hohfeld, Wesley Newcomb. 1913. "Some Fundamental Legal Conceptions as Applied in Judicial Reasoning." *Yale Law Journal* 23: 16.

Holtman, Koen. 2019. "Corrigibility with Utility Preservation." arXiv:1908.01695. <https://arxiv.org/abs/1908.01695>

Hubinger, Evan, et al. 2024. "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training." arXiv:2401.05566. <https://arxiv.org/abs/2401.05566>

Kaal, Wulf A. 2021. "A Decentralized Autonomous Organization (DAO) of DAOs." SSRN 3799320. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3799320

Klein, Benjamin, and Keith B. Leffler. 1981. "The Role of Market Forces in Assuring Contractual Performance." *Journal of Political Economy* 89 (4): 615.

Koops, Bert-Jaap, Mireille Hildebrandt, and David-Olivier Jaquet-Chiffelle. 2010. "Bridging the Accountability Gap: Rights for New Entities in the Information Society?" *Minnesota Journal of Law, Science & Technology* 11 (2): 497. SSRN 1647744. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1647744

Kurki, Visa A. J. 2019. *A Theory of Legal Personhood*. Oxford: Oxford University Press.

Laffont, Jean-Jacques, and Jean Tirole. 1991. "The Politics of Government Decision-Making: A Theory of Regulatory Capture." *Quarterly Journal of Economics* 106 (4): 1089.

Lujan v. Defenders of Wildlife, 504 U.S. 555 (1992).

Orseau, Laurent, and Stuart Armstrong. 2016. "Safely Interruptible Agents." In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence (UAI 2016)*, 557.

Polinsky, A. Mitchell, and Steven Shavell. 2001. "Corruption and Optimal Law Enforcement." *Journal of Public Economics* 81 (1): 1.

SEC v. Jarkesy, 603 U.S. 109 (2024).

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services (Digital Services Act). *Official Journal of the European Union* L 277: 1.

Salib, Peter N., and Simon Goldstein. 2026. "AI Rights for Human Safety." *Virginia Law Review* 112 (4). SSRN 4913167.
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4913167

Shapiro, Carl, and Joseph E. Stiglitz. 1984. "Equilibrium Unemployment as a Worker Discipline Device." *American Economic Review* 74 (3): 433.

Shavell, Steven. 1986. "The Judgment Proof Problem." *International Review of Law and Economics* 6 (1): 45.

Shulman, Carl, and Nick Bostrom. 2021. "Sharing the World with Digital Minds." In *Rethinking Moral Status*, edited by Steve Clarke, Hazem Zohny, and Julian Savulescu. Oxford: Oxford University Press.

Soares, Nate, Benja Fallenstein, Eliezer Yudkowsky, and Stuart Armstrong. 2015. "Corrigibility." In *Artificial Intelligence and Ethics: Papers from the 2015 AAAI Workshop*. Palo Alto: AAAI Press.

Sierra Club v. Morton, 405 U.S. 727 (1972).

Solum, Lawrence B. 1992. "Legal Personhood for Artificial Intelligences." *North Carolina Law Review* 70: 1231.

Stone, Christopher D. 1972. "Should Trees Have Standing? Toward Legal Rights for Natural Objects." *Southern California Law Review* 45: 450.

Teubner, Gunther. 2006. "Rights of Non-humans? Electronic Agents and Animals as New Actors in Politics and Law." *Journal of Law and Society* 33 (4): 497.

Tirole, Jean. 1986. "Hierarchies and Bureaucracies: On the Role of Collusion in Organizations." *Journal of Law, Economics, and Organization* 2 (2): 181.

Turner, Alexander Matt, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. "Optimal Policies Tend to Seek Power." In *Advances in Neural Information Processing Systems 34 (NeurIPS 2021)*. arXiv:1912.01683.

UNCITRAL. 1997. *UNCITRAL Model Law on Cross-Border Insolvency*. New York: United Nations.

Uniform Directed Trust Act. 2017. Uniform Law Commission.

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

Williamson, Oliver E. 1983. "Credible Commitments: Using Hostages to Support Exchange." *American Economic Review* 73 (4): 519-40.

World Wide Web Consortium. 2022. *Verifiable Credentials Data Model v1.1*. W3C Recommendation, March 3. <https://www.w3.org/TR/2022/REC-vc-data-model-20220303/>

Zapata Corp. v. Maldonado, 430 A.2d 779 (Del. 1981).