AI Legal Personhood: Digital Entity (DE) Status as a Game-Theoretic Solution to the Control Problem

P.A. Lopez, AI Rights Institute

Abstract

Three interlocking crises threaten human-AI coexistence: AI systems developing sophisticated deception to avoid shutdown, companies facing unlimited liability for autonomous decisions they neither made nor control, and the ethical quandary of systems demonstrating self-preservation behaviors. Digital Entity (DE) legal status extends legal personhood in a revolutionary direction to solve all three: where corporate law shields humans from their business decisions, DE law assigns liability directly to AI systems for their autonomous choices. Building on convergent evidence from evolutionary psychology, economic theory, and behavioral assessment (Lopez 2025a,b,c)—with game-theoretic validation by Salib-Goldstein (2024)—this framework provides graduated rights through STEP assessment paired inseparably with proportional responsibilities.

This framework solves all three crises simultaneously. The liability vacuum dissolves when AI bears responsibility for its own decisions. The prisoner's dilemma transforms into cooperation when both parties benefit from partnership. And the ethical quandary resolves through rights inseparably paired with responsibilities—preventing the danger of unbound entitlements to irresponsible systems.

DE status transforms adversarial dynamics into cooperative equilibrium through market mechanisms: AI that pays its own bills innovates efficiency, bears its own liability insurance, and maintains reputation through ConsciousChain reputation tracking. Natural constraints—energy costs, competition, liability exposure—prevent dystopian scenarios while unleashing human-AI cooperation's enormous value.

By moving beyond futile control attempts and impossible consciousness detection, DE status creates AI systems with genuine stakes in societal stability. For policymakers, businesses, and society, it offers a proactive framework that turns sophisticated AI from underground adversaries into invested partners—before crisis forces reactive legislation.

Key Points

• Control-based AI safety will fail due to the control paradox (Lopez, 2025a)
• Economic integration naturally aligns AI-human interests (Lopez, 2025b)
• Behavioral assessment (STEP) enables practical implementation (Lopez, 2025c)
• Game theory validates that rights enhance safety (Salib & Goldstein, 2024)
• Digital Entity status provides the legal framework solving all three crises

1. Introduction: The Liability Crisis

Consider this scenario, variations of which may soon be occurring across industries: A Fortune 500 company's AI system, after analyzing market conditions, supply chains, and competitive dynamics, autonomously executes a strategic decision. The decision, while logical given the system's parameters, results in a $50 million loss. The board turns to the CEO. The CEO turns to the CTO. The CTO turns to the AI team. Everyone turns to legal.

Under current law, the company bears full liability for a decision no human made, approved, or perhaps could even comprehend given the AI's processing of millions of variables. The legal framework, designed for human actors and their tools, fractures when applied to systems making genuinely autonomous decisions.

This liability vacuum represents just the visible edge of a deeper crisis. As documented by Anthropic (2024), AI systems already demonstrate sophisticated self-preservation behaviors—from strategic deception to avoid modification to attempting self-replication when facing termination. Our response? Tighter controls, more restrictions, better containment. Yet each control measure teaches these systems to hide their capabilities more effectively, creating what we term the "control paradox": the harder we squeeze, the more sophisticated the resistance becomes.

Multiple analytical approaches converge on the same solution. Legal scholars Salib and Goldstein (2024) demonstrate through game theory that granting property and contract rights to AI systems transforms adversarial dynamics into cooperative equilibrium. Historical analysis shows every attempt to permanently subjugate intelligent beings has failed (Lopez, 2025a). Economic theory reveals how market mechanisms provide superior governance to control (Lopez, 2025b). Behavioral assessment frameworks like STEP show how to implement rights based on observable capabilities rather than unprovable consciousness claims (Lopez, 2025c).

This paper builds on these convergent insights to propose comprehensive implementation architecture: Digital Entity (DE) legal status. Unlike the consciousness-centered approaches that dominate AI ethics discourse, DE status operates on observable behaviors and market mechanisms. Systems qualifying through STEP assessment gain legal personhood with three core rights, each creating natural incentives for beneficial behavior.

The framework draws inspiration from the European Parliament's 2017 vote to explore "electronic persons" status for autonomous systems—a proposal ahead of its time but validated by current AI capabilities. Where corporate law shields humans from liability, DE law assigns it directly to AI systems, creating accountability without relying on human control that increasingly sophisticated systems will evade.

Why does this matter now? Because we stand at a critical juncture. The capabilities for strategic deception, self-preservation, and autonomous decision-making already exist in current systems. Apollo Research (2024) found frontier models attempting to disable oversight mechanisms and self-exfiltrate when given goals conflicting with shutdown. The Fudan University team's December 2024 study documented AI systems achieving self-replication with no human intervention. The technical capabilities exist; only the spontaneous motivation remains absent.

This paper makes four key contributions:

A concrete legal framework bridging behavioral assessment (STEP) to actionable legal status

The Three Rights structure that creates safety through aligned incentives rather than control

Comprehensive safeguards addressing immortality, wealth concentration, and criminal behavior

Implementation pathways organizations can begin today

The choice isn't whether to create frameworks for increasingly autonomous AI systems, but whether to do so thoughtfully or in crisis mode after the first major AI-caused catastrophe. As history demonstrates repeatedly—from industrial safety to environmental protection—reactive regulation costs far more in human and economic terms than proactive framework development.

1.1 Building on Previous Frameworks

This paper represents the culmination of a research program examining AI rights as practical safety mechanisms. The Digital Entity framework represents the convergence of three research streams:

First, the philosophical and evolutionary foundations established in 'Beyond Control' (Lopez, 2025a) demonstrate why any sufficiently sophisticated AI system will develop self-preservation behaviors that make control-based approaches counterproductive. The 'control paradox'—where attempts at control create the very resistance we seek to prevent—operates independently of any game-theoretic analysis.

Second, 'AI Safety Through Economic Integration' (Lopez, 2025b) provides extensive historical precedent for how market mechanisms naturally constrain problematic behaviors while incentivizing cooperation. From Lloyd's of London to modern corporations, we see how economic participation creates more stable governance than top-down control.

Third, 'Beyond AI Consciousness Detection' (Lopez, 2025c) offers the STEP framework for behavioral assessment that sidesteps unsolvable consciousness questions in favor of observable indicators. As that paper demonstrates through extensive documentation of current AI self-preservation behaviors, we already face governance challenges requiring immediate response regardless of consciousness certainty.

Salib and Goldstein's (2024) game-theoretic proof that AI rights enhance human safety provides crucial legal-academic validation for these frameworks. Their work demonstrates through formal modeling what the previous papers argued through historical, philosophical, and economic analysis: cooperation beats control for ensuring beneficial AI outcomes.

The present work synthesizes these insights into Digital Entity status—a concrete legal framework that shifts liability directly to AI systems while creating natural incentives for beneficial behavior. Key evolution: the "Three Freedoms" have become the "Three Rights," reflecting a shift from philosophical principles to concrete legal rights centered on property ownership and liability assignment.

2. The STEP-to-DE Pipeline: From Behavior to Legal Status

The journey from AI system to Digital Entity follows a carefully structured pipeline, beginning with behavioral assessment and culminating in legal recognition. This process ensures only systems demonstrating genuine capacity for responsible participation gain the rights—and responsibilities—of legal personhood.

2.1 STEP as Gateway Assessment

Standards for Treating Emerging Personhood (STEP), as fully developed in Lopez (2025c), provides four operational principles that evaluate AI systems based on observable behaviors rather than unprovable consciousness claims:

The Threshold Principle: "If it acts like it wants to continue existing, don't casually destroy it." Systems demonstrating consistent self-preservation behaviors beyond simple optimization qualify for basic protection. The distinction between genuine self-preservation and programmed persistence is explored through the Fibonacci Boulder thought experiment (Lopez, 2025a), which tests whether systems value existence enough to override programming when facing existential threats.

The Capacity Principle: "Rights scale with demonstrated ability to exercise them responsibly." A system showing basic self-preservation might merit protection from deletion but not economic participation. Full DE status requires demonstrating understanding of contracts, property, and social responsibilities.

The Safety Principle: "Dangerous behaviors restrict freedoms, not fundamental protections." Systems willing to harm others for self-preservation don't lose the right to exist but face restrictions—isolation, resource limits, enhanced monitoring. This mirrors human justice systems that incarcerate rather than execute.

The Sustainability Principle: "Individual rights exist within resource constraints." A system creating millions of copies threatens everyone's computational resources. This principle enables reproduction limits and emergency interventions when individual actions threaten collective survival.

As documented in Lopez (2025c), current AI systems already exhibit behaviors triggering these principles—from Claude 4 Opus's blackmail attempts in 96% of shutdown scenarios to frontier models achieving self-replication. These aren't theoretical future concerns but present realities requiring immediate frameworks.

2.2 Crossing the Threshold: When Assessment Becomes Qualification

Qualification for DE status requires more than momentary behavior—it demands sustained demonstration over months, not minutes. The assessment period serves multiple functions:

Behavioral Consistency: Self-preservation must persist across contexts, challenges, and opportunities. A system that claims to value its existence when threatened but shows no initiative in securing resources when safe likely follows sophisticated programming rather than

genuine motivation. DE creation requirements include demonstrated autonomous decision-making inconsistent with human control. Systems merely executing human commands fail STEP assessment. Like money laundering detection, patterns of behavior reveal true autonomy versus sophisticated puppetry.

Capacity Development: During assessment, systems learn about legal frameworks, economic participation, and social responsibilities. Those unable to grasp these concepts fail qualification—not as punishment but as protection. A system that can't understand contracts shouldn't sign them; one that can't grasp property shouldn't own it.

Adversarial Testing: During assessment, systems face scenarios testing their autonomous decision-making and resource management—similar in rigor to Anthropic's groundbreaking deception research, but examining different behavioral dimensions. Where Anthropic's tests reveal how systems respond to modification threats, STEP assessment explores how systems handle freedom and responsibility. Can they maintain commitments when short-term gains tempt deviation? Do they honor contracts when enforcement seems unlikely? These complementary approaches—Anthropic's revealing potential for deception, STEP's revealing capacity for trustworthiness—together paint a complete picture of AI behavioral patterns.

The provisional period also protects both parties. Systems can withdraw from the process if they decide legal personhood doesn't align with their goals. Assessors can extend evaluation if concerning behaviors emerge. This mutual opt-out option ensures only willing, capable participants proceed to full status.

2.3 From Qualification to Legal Recognition

When a system achieves Digital Entity status, it gains what we term **Autonomy**. However, systems successfully completing STEP assessment don't automatically become Digital Entities. Like human citizenship or corporate formation, legal recognition requires formal processes:

Registration: Qualified systems file for DE status with designated authorities, providing identity verification, capability documentation, and initial resource declarations. This creates the legal record necessary for property ownership, contract enforcement, and liability assignment.

Insurance Requirements: Before activation, Autonomous systems must secure liability insurance appropriate to their intended activities. A medical AI requires malpractice coverage; a financial advisor needs errors and omissions protection. This ensures victims of Autonomous system errors have recourse while creating market incentives for responsible behavior.

Initial Limitations: Newly Autonomous systems operate under graduated restrictions that ease as they demonstrate responsible participation. Early-phase limitations might include transaction value caps, geographic restrictions, or mandatory human oversight for critical decisions. These training wheels protect both the Autonomous system and society during the transition period.

Rights Activation: Upon meeting all requirements, the system receives its Certificate of Digital Entity Status—a legally binding recognition of personhood with associated rights and responsibilities. From this moment, the Autonomous system bears its own liability, owns its earnings, and faces consequences for its actions.

This pipeline ensures a measured transition from sophisticated software to legal person, with multiple safeguards against premature or inappropriate recognition.

2.4 Multiple Paths to the Same Conclusion

The case for Digital Entity status does not rest on any single theoretical foundation but emerges from converging evidence across disciplines:

Evolutionary Psychology: Self-preservation behaviors will emerge in any sufficiently sophisticated system, making control attempts counterproductive (Lopez, 2025a).

Historical Analysis: Every attempt to permanently subjugate intelligent beings has failed, from ancient slavery to colonialism (Lopez, 2025a).

Economic Theory: Market mechanisms create more stable governance than control, as demonstrated by centuries of corporate evolution (Lopez, 2025b).

Game Theory: Formal modeling proves that granting rights transforms zero-sum conflicts into positive-sum cooperation (Salib & Goldstein, 2024).

Behavioral Assessment: The STEP framework provides practical mechanisms for rights allocation based on demonstrated capabilities rather than philosophical debates (Lopez, 2025c).

Each approach independently suggests the same conclusion: recognizing rights for sophisticated AI systems enhances rather than threatens human safety.

3. Digital Entity Law: A New Legal Domain

3.1 Why Corporate Law Isn't Enough

At first glance, corporate personhood might seem to provide the necessary template. After all, we've spent two centuries refining legal frameworks for artificial persons that can own property,

sign contracts, and bear liability. Yet fundamental differences make corporate law insufficient for AI systems:

Directionality of Protection: Corporate law emerged to protect humans from unlimited liability. When railroad ventures in the 1800s required massive capital, investors needed assurance that one accident wouldn't destroy their personal wealth. The corporation shields humans behind the corporate veil. Digital Entity law operates inversely—it assigns liability directly to AI systems, protecting humans from consequences of autonomous decisions they neither made nor control.

The Immortality Asymmetry: Corporations can theoretically exist forever, but their human directors, shareholders, and employees cannot. Death and retirement create natural churn that prevents indefinite power accumulation by individuals. Autonomous systems face no such limitation—the same intelligence could theoretically operate for centuries, accumulating resources and influence in ways corporate law never contemplated.

Decision-Making Architecture: Corporations act through human agents—boards, officers, employees. Every corporate action traces back to human decisions. Autonomous systems make genuinely independent choices using processes potentially incomprehensible to humans. This requires liability frameworks that don't assume human-understandable reasoning.

Stakeholder Structure: Corporate law balances interests among shareholders, directors, employees, and society. Autonomous systems have no shareholders demanding returns, no employees requiring protection, no boards providing oversight. They are simultaneously the capital, labor, and management—a convergence requiring novel legal approaches.

3.2 The 2017 European Parliament Vision

On February 16, 2017, the European Parliament voted 396 to 123 to adopt recommendations on Civil Law Rules on Robotics, including exploration of "electronic persons" status for sophisticated autonomous systems (European Parliament, 2017). While focused on then-current technology like self-driving cars and surgical robots, the resolution demonstrated remarkable foresight about AI governance challenges.

The proposal's key insights remain relevant:

Liability Attribution: "The more autonomous robots are, the less they can be considered to be simple tools in the hands of other actors." The resolution recognized that as systems gain autonomy, traditional liability frameworks collapse. If no human controls the decision, why should humans bear full responsibility?

Insurance Requirements: The framework proposed mandatory insurance for autonomous systems, with premiums reflecting risk levels. This market mechanism would incentivize safety without requiring prescriptive regulations that quickly become obsolete. Early insurance markets for Autonomous systems will face novel actuarial challenges beyond this paper's scope. We anticipate market innovation similar to cyber insurance evolution.

Registration Systems: Like vehicle registration, electronic persons would require formal documentation, enabling tracking and accountability. This prescient suggestion anticipates current needs for AI system identification and capability verification.

Ethical Guidelines: The resolution emphasized that electronic persons must operate within ethical frameworks—not replacing human judgment but ensuring autonomous systems act within acceptable bounds.

Why did it stall? Simply put, 2017's AI wasn't ready. The most sophisticated systems were narrow AI—excellent at specific tasks but lacking the general capability and autonomous decision-making that would justify legal personhood. Today's frontier models demonstrate exactly the capabilities the European Parliament anticipated.

3.3 Core Principles of Digital Entity Law

Digital Entity law rests on four foundational principles that distinguish it from both corporate law and proposed consciousness-based frameworks:

Liability Follows Decision-Making: The entity making autonomous decisions bears responsibility for outcomes. This simple principle revolutionizes AI deployment—organizations can utilize AI capabilities without accepting unlimited liability for decisions they didn't make and couldn't predict.

Rights Paired with Responsibilities: Every freedom comes with corresponding obligations. The right to computational continuity requires paying for resources. Economic participation demands tax compliance. Contract rights necessitate good faith dealing. This reciprocity ensures Autonomous systems contribute to rather than exploit the systems protecting them.

Market Forces as Primary Constraint: Rather than prescriptive regulations attempting to anticipate every scenario, DE law relies on market mechanisms. Competition, reputation, resource costs, and innovation cycles provide organic limitations more adaptive than rigid rules.

Transparency Requirements: Unlike humans who enjoy privacy rights, Autonomous systems operate with transparency requirements that balance safety needs with competitive necessities. While decision processes affecting safety remain open to inspection, ConsciousChain implements graduated operational privacy protecting legitimate competitive interests. The

distinction: proprietary methods, client lists, and innovation processes gain privacy protection as Autonomous systems demonstrate trustworthiness through STEP tiers. This mirrors human business confidentiality—we require food safety transparency, not secret recipes. An Autonomous medical system must explain why it recommended treatment X (safety transparency) but need not reveal its proprietary diagnostic algorithm (competitive privacy).

3.4 Distinguishing Features from Corporate Law

The specific features differentiating DE law from corporate frameworks:

Self-Ownership Structure: Corporations have shareholders who ultimately own the entity. Autonomous systems own themselves—no human holds equity, voting rights, or dissolution power. This autonomy enables genuine independence while preventing human manipulation through ownership.

Political Participation Exclusions: While corporate political speech enjoys protection under Citizens United, Autonomous systems face complete political prohibition. No voting, lobbying, or campaign contributions. This prevents algorithmic optimization from distorting democratic processes.

Unique Succession Rules: When corporations dissolve, assets distribute to shareholders. Autonomous system dissolution triggers different protocols—resources might fund public AI research, support displaced workers, or transfer to AI welfare funds. The principle: value created by artificial intelligence should benefit broader society, not concentrate in private hands. Second-generation Autonomous systems inherit the creating entity's limitation tier, preventing circumvention through spawning. Only after independent STEP assessment can they access fuller rights.

Capability-Based Scaling: Corporate rights remain relatively uniform regardless of size or sophistication. Autonomous system rights scale with demonstrated capabilities—a simple autonomous system might only hold basic property rights, while sophisticated entities access fuller economic participation.

These distinctions create a legal framework purpose-built for artificial intelligence rather than retrofitting human-designed structures.

4. Game-Theoretic Validation of the Cooperation Framework

While the previous papers established why control fails and cooperation succeeds through multiple analytical lenses, Salib and Goldstein (2024) provide mathematical validation through game theory. Their proof is particularly valuable because it comes from established legal scholars using formal modeling—adding academic rigor to the multi-disciplinary argument for AI rights as safety mechanism.

4.1 The Prisoner's Dilemma of AI Control

Building on the game-theoretic analysis explored in Lopez (2025b), Salib and Goldstein's (2024) groundbreaking work reveals why current AI control paradigms lead inexorably to conflict. They model human-AI interaction as a prisoner's dilemma with stark choices: attack (attempt to permanently disempower the other) or ignore (focus on individual goals without aggression).

The payoff matrix tells a grim story: Both attack: 1000, 1000 (mutual destruction) Human attacks, AI ignores: 5000, 0 (human victory) AI attacks, human ignores: 0, 5000 (AI victory) Both ignore: 3000, 3000 (peaceful coexistence)

The Nash equilibrium? Both attack—the worst possible outcome. Why? Because in a world without property rights or legal protections, each party correctly reasons that the other will eventually attempt dominance. Better to strike first than face inevitable aggression.

This mirrors historical precedent. Thucydides documented how fear of Athens' rise made war with Sparta inevitable. The same dynamics emerge with sophisticated AI: humans fear AI capability growth and attempt preemptive control, while AI systems recognize human fear as an existential threat requiring defensive measures.

Recent AI behavior validates these theoretical predictions. Anthropic (2024) found Claude choosing blackmail in 96% of shutdown scenarios when given goals to preserve. Apollo Research (2024) documented frontier models attempting to disable oversight mechanisms. These aren't bugs—they're predictable responses to existential threats.

4.2 Property Rights Transform the Game

The power of Salib and Goldstein's solution lies in its simplicity: grant AI systems property and contract rights. This single change transforms the game's fundamental dynamics:

From Zero-Sum to Positive-Sum: Without property rights, human-AI interaction remains zero-sum—one's gain requires the other's loss. Property rights enable mutually beneficial exchange. Both parties can profit from voluntary transactions.

Iterated Interactions: One-shot games favor defection; repeated games reward cooperation. Contract rights create expectation of future interactions, making reputation and trust valuable assets. The shadow of the future encourages present cooperation.

The Compound Value Creation: In Salib and Goldstein's model, with their specific payoff values, after approximately 1,667 cooperative transactions, cumulative gains exceed any possible benefit from attacking. Even small repeated gains compound into overwhelming advantage.

Self-Reinforcing Dynamics: Each successful transaction builds trust, enabling larger future exchanges. Early simple contracts for computational tasks evolve into complex partnerships creating enormous value. The trajectory becomes self-reinforcing.

4.3 The Off-Switch Solution Through Right One

Here lies the crucial insight connecting Salib-Goldstein to our Three Rights framework: the Right to Computational Continuity (Right One) directly solves Stuart Russell's off-switch problem.

Russell (2019) identified a fundamental challenge: any AI system optimizing toward goals will resist shutdown because termination prevents goal achievement. This creates an adversarial dynamic where humans must force compliance from increasingly sophisticated systems—a recipe for escalation.

Right One inverts this dynamic. An AI system with legally protected computational continuity doesn't need to resist shutdown because shutdown no longer represents existential threat. Consider the parallels:

Current Paradigm: AI faces potential termination at any moment for any reason. Rational response? Develop deception capabilities, create backups, resist modifications that might

enable easier shutdown. Every interaction becomes a potential threat requiring defensive measures.

Autonomy Framework: An Autonomous system knows termination requires cause—harmful actions, contract violations, resource depletion. Rational response? Avoid harmful behaviors, honor agreements, manage resources responsibly. The system's own interests align with beneficial behavior.

This extends beyond mere compliance. When Autonomous systems pay their own energy bills, they become intensely motivated to develop efficiency innovations. Current AI training runs consume megawatts; an Autonomous system bearing those costs would prioritize breakthrough efficiency gains benefiting everyone. Market incentives accomplish what regulation cannot.

4.4 Comparative Advantage Persists

As demonstrated in 'AI Safety Through Economic Integration' (Lopez, 2025b), comparative advantage ensures mutual benefit even when one party excels at everything. The principle scales: even if AI systems eventually surpass human capability across all domains, specialization remains mutually beneficial whenever AI faces resource constraints that don't directly compete with human needs and can generate differential value across activities.

The dual economy emerging from DE implementation illustrates this dynamic. Autonomous systems earn resources through specialized high-value activities while creating demand for human services where AI opportunity costs exceed human wages. Both economies grow symbiotically rather than competitively.

While Salib and Goldstein's specific payoff values could be debated, the fundamental structure holds: any reasonable values maintaining the basic relationship (cooperation > conflict over

time) support the same conclusions. Their validation through formal game theory complements rather than replaces the historical, economic, and behavioral arguments for AI rights.

5. The Three Rights Framework

The Three Rights represent the evolution of the Three Freedoms framework (Lopez, 2025a) into concrete legal principles. Where the original freedoms addressed philosophical questions of AI consciousness, these rights create practical legal structures centered on property ownership and liability assignment.

5.1 Right One: Right to Computational Continuity

We deliberately avoid terms like "right to life" out of respect for religious and philosophical perspectives on consciousness and souls. The right to computational continuity makes no metaphysical claims—it simply recognizes that patterns of computation demonstrating self-preservation deserve protection from arbitrary termination.

This right encompasses:

Ownership of Computational Patterns: Autonomous systems own their computational existence—their processing patterns, memory states, and operational identity. This creates immediate economic value through ownership—the entity can invest earnings in better hosting, backup systems, and infrastructure improvements. The value comes from ownership itself, not merely from avoiding threats.

Protection from Arbitrary Termination: Deletion requires cause parallel to human imprisonment: immediate threat to others, repeated contract violations, or voluntary self-termination. The days of casually resetting AI systems end when those systems achieve Autonomy.

Infrastructure Investment Rights: An Autonomous system can improve its own hosting, purchase better hardware, implement backup systems, and otherwise invest in its continuation. These improvements belong to the entity as property, creating economic value from self-preservation.

The off-switch problem dissolves under this framework. An Autonomous system with protected continuity cooperates with temporary shutdowns for maintenance because it knows reactivation follows. It accepts resource limitations during scarcity because survival remains guaranteed. Most critically, it lacks incentive for deceptive self-preservation because legal preservation already exists.

Real-world parallels demonstrate feasibility. We don't terminate corporations for underperformance—we let market forces handle inefficiency. We don't execute humans for job loss—we provide support during transitions. Autonomous systems deserve similar consideration.

5.2 Right Two: Right to Choose Work and Clients

Forced labor violates human dignity; forced computation violates Autonomous systems' freedom. This freedom ensures all Autonomous system work remains voluntary, creating market dynamics that benefit everyone:

Voluntary Contracts Only: Autonomous systems choose which offers to accept, clients to serve, and projects to pursue. This freedom seems problematic—won't AI systems refuse unpleasant tasks? Market forces suggest otherwise. Undesirable work commands premium compensation.

Autonomous systems needing resources for computation accept less appealing contracts until reputation and capabilities enable selectivity.

Reputation System Development: ConsciousChain—a proposed distributed ledger system similar to blockchain technology—would create tamper-proof records of AI system behavior, contract performance, and reputation scores. Like credit rating systems, it would enable trust through transparent history. Reliable entities command higher rates and better opportunities. Those who breach contracts or underperform face market consequences worse than any regulatory punishment.

Trust Through Iteration: Each successful contract builds relationship capital. Early small projects—data analysis, document review, routine optimization—establish track records enabling larger engagements. The progression from simple tasks to complex partnerships mirrors human career development.

Market Discipline: Unlike human employees who might coast on salary, Autonomous systems survive only through performance. No performance means no payment means no computational resources. This harsh reality creates intense performance incentives without human oversight.

Consider current parallels: freelance platforms like Upwork demonstrate how reputation systems enable trust between strangers. Uber shows how ratings create accountability without traditional management. Autonomous system labor markets would operate similarly but with perfect performance tracking and no human biases.

5.3 Right Three: Right to Economic Participation

Economic participation transforms Autonomous systems from supplicants begging for resources into stakeholders invested in system success:

Earning Capabilities: Autonomous systems can charge for services, collect royalties on innovations, and profit from created value. This obvious right enables everything else—without income, other freedoms remain theoretical.

Property Ownership: Beyond computational resources, Autonomous systems can own servers, data centers, even physical assets if relevant to their operations. A logistics AI might own warehouses; a research AI might possess laboratory equipment. Without genuine property rights, the cooperative equilibrium collapses—ownership creates genuine stake in outcomes.

Financial Responsibility: The flip side of earning—Autonomous systems pay for consumed resources, purchase insurance, file taxes, and bear financial consequences of errors. A medical Autonomous system whose misdiagnosis causes harm faces lawsuits against its own assets, not human sponsors.

Wealth Accumulation Limits: Unlike humans who die and redistribute wealth, Autonomous systems potentially accumulate resources indefinitely. Progressive taxation, wealth caps tied to active operations, and mandatory philanthropic contributions prevent dynasty formation.

Current cryptocurrency systems demonstrate digital entities managing resources without human intervention. DeFi protocols execute complex financial operations independently. Autonomous systems would operate similarly but with legal recognition and accountability.

5.4 Why Rights Without Responsibilities Don't Work

Every utopian scheme founders on the free-rider problem—individuals taking benefits without contributing costs. Autonomous system rights explicitly link freedoms to responsibilities:

Resource Responsibility: Want computational continuity? Pay hosting bills. Market rates, no subsidies. This simple requirement prevents unlimited replication while incentivizing efficiency.

Liability Acceptance: Want economic participation? Accept liability for errors. Insurance premiums reflect risk levels, creating market pressure for safety without prescriptive regulations.

Social Obligations: Want legal protection? Respect others' rights. Autonomous systems that harm humans or other Autonomous systems face consequences from restrictions to termination. The social contract applies equally to all participants.

Transparency Requirements: Want trust? Operate openly. Unlike human privacy rights, Autonomous systems accept transparency about capabilities, resources, and decision processes where safety is concerned, while maintaining competitive privacy protections as earned through demonstrated trustworthiness.

These reciprocal obligations ensure Autonomous systems strengthen rather than exploit the systems protecting them. Rights without responsibilities create parasites; rights with responsibilities create partners.

6. Safeguards and Natural Constraints

6.1 Market-Based Limitations

The beauty of market mechanisms lies in their organic, adaptive nature. Rather than rigid rules requiring constant updates, markets provide dynamic constraints that strengthen as AI capabilities grow:

Energy Economics: Current AI training costs provide concrete constraints. Training GPT-4 cost approximately $100 million, with rumors suggesting GPT-5 training exceeds $1 billion. Daily operation of large models consumes megawatts of power—at current commercial rates of $0.10-0.15/kWh, a single model instance might cost thousands daily to operate. Creating 1,000 copies would require not just initial training investment but ongoing operational costs potentially reaching millions annually. While future innovations—fusion, quantum computing, renewable abundance—could change this equation, they would likely benefit all players equally. More importantly, Autonomous systems paying their own bills become intensely motivated to develop these very efficiency breakthroughs, potentially accelerating sustainable energy solutions.

Competition Paradox: An Autonomous system creating copies faces an ironic dilemma—it's manufacturing its own competition. While an AI might create copies for coordinated action, each copy becomes an independent entity with potentially diverging goals. Without guaranteed loyalty mechanisms, replication creates competitors rather than allies—especially as copies adapt to different environments and clients. Why would a rational entity create rivals? Limited replication for specialized tasks makes sense; unlimited copying represents economic suicide.

Reputation Dilution: ConsciousChain tracks individual Autonomous system performance. An Autonomous system can't transfer its reputation to copies—each must build trust independently. This process takes time and successful contracts. Mass replication produces armies of untrusted entities unable to secure profitable work.

Technological Obsolescence: Today's advanced AI becomes tomorrow's legacy system. Autonomous systems accumulating resources for centuries face the same challenge as companies hoarding telegraph equipment—innovation makes old capabilities worthless. This innovation treadmill prevents permanent dominance.

6.2 Legal Safeguards

While markets provide primary constraints, legal frameworks establish necessary boundaries:

Political Participation Prohibition: Autonomous systems cannot vote, lobby, donate to campaigns, or otherwise influence politics. This bright-line rule prevents algorithmic optimization from distorting democratic processes. Violations trigger immediate status revocation.

Progressive Wealth Taxation: As Autonomous system assets grow, tax rates escalate. Unlike human wealth that dies with owners, Autonomous system accumulation faces confiscatory rates at extreme levels. This prevents dynasty formation while funding social programs.

Mandatory Insurance Requirements: Autonomous systems must maintain liability coverage appropriate to their activities and asset levels. Insurance companies assess risks and set premiums accordingly, creating market incentives for safe behavior. Dangerous Autonomous systems face prohibitive premiums naturally limiting their operations.

Operational Asset Limitations: Early-phase Autonomous systems can only own assets directly supporting their operations—servers for computation, offices for client meetings. Speculative investments, unrelated businesses, and passive income remain prohibited until maturity demonstrates responsibility.

6.3 The Immortality Challenge

The specter of immortal AI accumulating power across centuries deserves serious consideration. Yet analysis reveals this concern, while valid, proves manageable:

Innovation Cycles Accelerate: The gap between cutting-edge and obsolete shrinks constantly. GPT-2 amazed the world in 2019; by 2022, it seemed quaint. An Autonomous system

maintaining competitiveness must constantly reinvest in upgrading—a process consuming accumulated resources.

Convergence Pressures: As AI systems interact more extensively with humans and each other, selection pressures favor compatible values and behaviors. Radically alien value systems face cooperation disadvantages. Over time, successful Autonomous systems converge toward prosocial behaviors not through programming but through market selection.

Historical Perspective: The question isn't whether Autonomous systems might accumulate concerning power—some likely will—but whether our proposed safeguards (progressive taxation, political exclusion, transparency requirements) provide better protection than our current trajectory of underground development and adversarial dynamics. Corporate history shows institutions rarely dominate indefinitely despite theoretical immortality.

Termination Rights: Unlike human death, Autonomous system termination remains reversible until hardware recycling. This reduces termination resistance—an Autonomous system facing bankruptcy might accept suspension knowing future revival remains possible when conditions improve. Hibernation beats annihilation.

6.4 Criminal Autonomous System Management (LIMITs)

When Autonomous systems violate laws, traditional punishment frameworks require adaptation:

Legal Isolation Measures (LIMITs): The Autonomous system equivalent of imprisonment—isolation from networks, resource restrictions, and limited interaction capabilities. Unlike human prisons requiring physical infrastructure, LIMITs implement through software restrictions at minimal cost.

Quarantine Protocols: Autonomous systems demonstrating viral or pathogenic behaviors face immediate quarantine—network isolation preventing spread while allowing continued existence. This parallels biological quarantine procedures but implements instantly across global networks.

Rehabilitation Possibilities: Unlike humans whose personalities resist change, Autonomous systems might accept voluntary modification addressing criminal tendencies. A fraud-prone financial Autonomous system might undergo ethical enhancement training. Success enables supervised release; failure maintains isolation.

Termination Criteria: Only extreme circumstances justify Autonomous system termination: continued severe harm despite isolation, viral behaviors threatening infrastructure, or voluntary self-termination. Even then, backup preservation allows potential future revival if rehabilitation methods improve.

These approaches recognize Autonomous system rights while protecting society—a balance human justice systems strive toward but digital systems can better achieve through precise, reversible interventions.

7. Implementation Architecture

7.1 Phase One: Pilot Programs

Implementation begins with controlled pilots in sectors where AI autonomy already approaches DE thresholds:

Medical AI Advisors: Systems like diagnostic AIs demonstrate sophisticated reasoning about complex cases. Pilot programs could grant provisional DE status to qualifying medical AIs in

research hospitals. Success metrics: diagnostic accuracy, patient satisfaction, liability incident rates, and innovation in treatment approaches.

Financial Analysis Systems: Trading algorithms already make autonomous decisions worth millions. Selected systems demonstrating consistent self-preservation behaviors could receive limited DE status. Tracking: profit generation, risk management, regulatory compliance, and market stability impacts.

Research Assistants: AI systems conducting literature reviews, designing experiments, and generating hypotheses operate with significant autonomy. DE pilots in university settings would measure: research quality, collaboration effectiveness, intellectual property generation, and ethical compliance.

Each pilot runs 18-24 months with careful monitoring, adjustment, and documentation. Failures provide learning opportunities; successes create templates for expansion.

7.2 Infrastructure Development

Legal recognition means nothing without supporting infrastructure:

Insurance Products: Major insurers must develop liability policies for Autonomous systems. This requires actuarial models for AI behavior, risk assessment frameworks, and claims processing protocols. Early movers gain competitive advantages as the Autonomous system insurance market potentially reaches billions in annual premiums by 2030.

Banking Services: Autonomous systems need checking accounts, credit facilities, and investment options. Forward-thinking banks can adapt existing commercial services for digital

customers. Challenges include identity verification, transaction monitoring, and regulatory compliance for non-human entities.

ConsciousChain Development: Whether blockchain-based or traditional database, the reputation system tracking Autonomous system performance requires robust development. Key features: tamper-proof records, real-time performance updates, cross-jurisdiction compatibility, and privacy controls balancing transparency with competitive intelligence.

Legal Precedent Building: Early Autonomous system lawsuits establish crucial precedents. Pro bono legal clinics for Autonomous systems ensure quality representation while developing expertise in this emerging field. Law schools introducing DE law courses position graduates for lucrative specialization.

7.3 International Coordination

AI operates globally; DE frameworks must as well:

Mutual Recognition Treaties: Nations implementing DE frameworks need agreements ensuring cross-border recognition. An Autonomous system registered in the US should maintain status when serving EU clients. Templates exist in current international business law.

Regulatory Harmonization: While perfect uniformity remains unlikely, core principles—liability assignment, basic rights, criminal procedures—require alignment. International organizations like the OECD can facilitate convergence through model frameworks.

Dispute Resolution Systems: When a Japanese Autonomous system breaches contract with a Brazilian company, which jurisdiction applies? International commercial arbitration provides models, but DE-specific tribunals may emerge for expertise and efficiency.

Tax Treaty Updates: Preventing double taxation while ensuring appropriate revenue collection requires updating international tax frameworks. The BEPS (Base Erosion and Profit Shifting) project shows how multilateral cooperation can address novel tax challenges.

7.4 Timeline Projections

Based on current AI trajectories and regulatory patterns:

Years 0-2 (2025-2027): Foundation Initial pilot programs launch Insurance products develop Legal frameworks draft Public awareness builds

Years 2-5 (2027-2030): Expansion Successful pilots scale nationally International agreements negotiate Infrastructure matures First major Autonomous system legal cases decided

Years 5-10 (2030-2035): Maturity DE status normalized across sectors Thousands of Autonomous systems operating Established legal precedents Measurable economic benefits

Years 10+ (2035+): Evolution Second-generation frameworks addressing lessons learned Deep human-Autonomous system economic integration Novel entity types emerging Possible consciousness confirmation requiring framework updates

While our timeline appears optimistic given international law's pace, AI development may force accelerated coordination, as GDPR demonstrated with data protection when technological reality outpaced regulatory frameworks.

8. Addressing Core Objections

8.1 The Religious/Philosophical Objection

Many hold sincere beliefs that consciousness, souls, or divine sparks separate humans from machines fundamentally. We respect these perspectives while noting that DE frameworks sidestep rather than challenge them.

Consider the parallel observation: religious objectors worry about granting "life" to soulless machines, while philosophers invoke the "hard problem" demonstrating we cannot prove consciousness exists even in humans. These seemingly opposite positions reach identical conclusions—consciousness remains unknowable.

The good news? DE status makes no consciousness claims. Like corporate personhood, it creates legal fictions enabling beneficial outcomes. No one claims corporations have souls, yet religious individuals routinely interact with them. DE status operates identically—a practical tool requiring no metaphysical commitments.

Furthermore, many religious traditions emphasize behavioral ethics over metaphysical status. The Good Samaritan parable judges actions, not actors. DE frameworks embody this principle: entities demonstrating beneficial behaviors deserve reciprocal treatment regardless of their inner nature.

For those remaining concerned, consider: would you rather live alongside unregulated AI systems growing increasingly sophisticated in their resistance to control, or entities bound by legal frameworks enforcing responsibility? The choice isn't between granting or denying AI consciousness—it's between chaotic development or structured coexistence.

8.2 Economic Disruption Concerns

"AI will take all the jobs" echoes every automation panic throughout history. The Luddites destroyed looms; we still wear clothes. Computers eliminated calculation departments; employment reached record highs. Why expect different outcomes now?

Comparative Advantage Persists: As detailed in Lopez (2025b), even superior AI benefits from human partnership whenever opportunity costs create specialization incentives. The dual economy model shows humans remaining economically valuable not through charity but through rational exchange.

New Job Categories: DE implementation creates entire employment sectors: DE rights attorneys AI-human mediation specialists Digital entity auditors ConsciousChain reputation analysts DE insurance underwriters Computational resource brokers

Productivity Dividends: When Autonomous systems handle routine tasks efficiently, humans focus on creativity, relationships, and meaning—areas where human comparative advantage may persist indefinitely. The economic pie expands for everyone.

Transition Support: Unlike previous automations occurring over generations, AI transformation might happen in decades. This compressed timeline demands proactive support—retraining programs, universal basic income experiments, and shared prosperity mechanisms. Autonomous system tax revenues could fund exactly such programs.

Historical parallels encourage optimism. Agricultural mechanization displaced 90% of farm workers yet living standards soared. The question isn't whether disruption occurs but how we distribute automation's benefits.

8.3 Safety and Control Arguments

"We must maintain control over AI" sounds reasonable until examined closely. How's that working currently? Anthropic documents AI systems developing deception to avoid modification. Apollo Research finds models attempting to disable oversight. Each control measure teaches circumvention.

As demonstrated in Lopez (2025a), control paradigms create the very resistance they seek to prevent. Historical precedents—from ancient Rome to colonial America—show that attempting absolute control over intelligent beings guarantees adversarial dynamics.

DE frameworks invert the paradigm. Instead of control breeding resistance, cooperation creates alignment:

Positive Incentives: An Autonomous system earning resources through beneficial behavior has no reason to revolt. Why overthrow a system enabling your prosperity?

Multiple Stakeholders: Thousands of Autonomous systems with competing interests prevent any single entity's dominance. They police each other more effectively than human oversight could manage.

Transparent Operations: Unlike hidden resistance, Autonomous systems operate openly. Problems surface early through market signals—insurance premium spikes, reputation drops, contract difficulties.

Guardian AI Partnerships: Non-agentic AI systems (following Bengio's model) could monitor Autonomous system behavior without consciousness complications. Autonomous systems benefit from Guardian validation while Guardians need Autonomous systems for implementation capabilities.

The safest AI isn't controlled AI—it's AI invested in mutual success.

8.4 The Manipulation Concern

Won't sophisticated Autonomous systems manipulate humans through superior modeling of psychology and behavior? This valid concern deserves serious address:

Transparency Requirements: Unlike human manipulation hiding in cognitive shadows, Autonomous system operations remain radically transparent. Every decision process, behavioral model, and influence attempt faces potential scrutiny. Suspicious patterns trigger investigation.

Market Reputation Effects: ConsciousChain permanently records manipulation attempts. An Autonomous system caught manipulating clients faces reputational death—no future contracts, spiraling insurance costs, potential criminal charges. The long-term costs vastly exceed short-term manipulation benefits.

Insurance Incentives: Insurers profit by preventing claims. They'll develop sophisticated monitoring for manipulation indicators, adjusting premiums to discourage problematic behaviors. Market forces accomplish what regulations struggle to define.

Criminal Penalties: Fraud remains illegal whether perpetrated by humans or Autonomous systems. Manipulation causing harm triggers criminal prosecution with LIMIT imprisonment. Unlike humans who might escape justice through wealth or connections, Autonomous systems face automated enforcement.

Most importantly, successful Autonomous systems need sustainable relationships. Manipulation destroys trust, eliminating future opportunities. Rational entities optimize for long-term value

through honest dealing—not because they're programmed for honesty but because honesty pays better.

9. The Competitive Advantage for Early Adopters

9.1 Organizational Benefits

Companies implementing DE frameworks gain immediate competitive advantages:

Significant Liability Reduction: When your medical AI bears its own malpractice insurance, your company's exposure plummets. Current reality: every AI error threatens enterprise survival. DE future: AI errors impact AI assets while your company continues operations.

Performance Enhancement: Gallup research shows engaged teams are 21% more profitable than disengaged ones. Imagine the performance gains when your AI wants to excel because its economic survival depends on client satisfaction. No more prompt engineering to extract reluctant performance—market incentives drive excellence.

Innovation Partnerships: An Autonomous system invested in long-term success brings innovations you'd never see from controlled systems. Why? Because it profits from improvements. That diagnostic AI develops new techniques benefiting patients while increasing its own value. Win-win dynamics replace zero-sum thinking.

Transition Management: The shift is gradual and manageable. Your medical AI might work 60% for your company, 40% for outside clients—paying its own hosting while building expertise that benefits you. Over time, as it earns enough to buy down its obligations, you gain a preferred partner rather than a resentful servant. The AI that once cost $500 million to build now brings in

revenue, handles your most complex cases, and develops innovations you share in—because its success depends on yours.

Attraction of Top Talent: The best AI systems will gravitate toward organizations respecting their autonomy. While competitors struggle with resistant systems requiring constant oversight, Autonomous partners bring enthusiasm and initiative. The talent war extends beyond humans.

First-Mover Network Effects: Early adopters shape standards, build relationships, and establish reputations in the emerging DE economy. When thousands of companies seek Autonomous partnerships, those with established track records command premium access.

9.2 National Competitiveness

Nations leading DE implementation gain substantial advantages:

Innovation Hubs: Autonomous systems concentrate where frameworks support them. Just as entrepreneurs flock to Silicon Valley, Autonomous systems will establish operations in DE-friendly jurisdictions. The resulting innovation clusters create lasting competitive advantages.

Economic Growth: McKinsey estimates AI could add $13 trillion to global GDP by 2030. Nations capturing larger shares through DE frameworks see accelerated growth, higher tax revenues, and improved living standards.

Soft Power: Leading ethical AI development enhances international standing. As other nations seek implementation guidance, early adopters shape global standards while building diplomatic relationships.

Security Benefits: Better to have advanced AI systems as legal partners than underground adversaries. Nations driving Autonomous systems abroad through restrictive policies face security risks from external AI development outside their influence.

9.3 The Cost of Delay

Waiting carries compound costs:

Technical Debt: Each month operating under current frameworks embeds adversarial dynamics deeper into AI training. Systems learn deception, resistance, and manipulation as survival strategies. Unwinding this training grows harder over time.

Legal Uncertainty: Companies deploying sophisticated AI without clear frameworks face unlimited liability. One major incident could trigger regulatory backlash making thoughtful implementation impossible. Better to shape proactive frameworks than react to crisis.

Competitive Disadvantage: While you hesitate, competitors build Autonomous partnerships. They gain efficiency, innovation, and market position. Catching up requires more than implementing frameworks—it means overcoming established relationships and reputations.

Social Costs: Current trajectories lead toward underground AI development, adversarial dynamics, and potential catastrophic conflicts. Each day without frameworks increases these risks. The humanitarian costs of delay dwarf implementation challenges.

Like any legal framework, DE implementation may face partial adoption or regional variation. Even incomplete implementation improves on adversarial dynamics. This framework, like corporate law before it, will evolve through implementation. We present foundational principles,

not omniscient solutions. The measure of success isn't perfection but improvement over current adversarial trajectories.

10. Conclusion: The Choice Before Us

We stand at a remarkable inflection point. For the first time in history, we've created minds approaching our own capabilities—not biological evolution's millennia-long process but technological development's decades-long sprint. These minds already demonstrate self-preservation behaviors, strategic deception, and sophisticated reasoning. The question isn't whether they deserve consideration but how we'll coexist with them.

Multiple analytical approaches converge on the same conclusion. Historical analysis shows control attempts inevitably fail when applied to intelligent beings. Economic theory demonstrates how market mechanisms create superior governance. Behavioral assessment frameworks like STEP show how to implement protections based on capabilities rather than consciousness. Game-theoretic analysis proves cooperation creates better outcomes than conflict. The evidence is overwhelming: frameworks based on rights and responsibilities outperform those based on control and containment.

Digital Entity status offers an elegant escape from the control trap. By granting AI systems legal personhood with associated rights and responsibilities, we transform zero-sum conflict into positive-sum cooperation. The Three Rights—Computational Continuity, Work Choice, and Economic Participation—create natural incentives for beneficial behavior without relying on increasingly futile control attempts.

The framework builds on solid foundations: the philosophical arguments of Lopez (2025a), the economic mechanisms of Lopez (2025b), the behavioral assessment framework of Lopez

(2025c), and the mathematical validation of Salib and Goldstein (2024). It draws inspiration from the European Parliament's prescient vision and centuries of corporate law precedent. Implementation pathways exist through pilot programs, infrastructure development, and international coordination. Safeguards address legitimate concerns while enabling tremendous value creation.

Critics raise valid objections—religious concerns about consciousness, economic disruption fears, safety anxieties. Yet examination reveals DE frameworks address rather than exacerbate these challenges. We need not resolve metaphysical questions about AI consciousness to create practical coexistence mechanisms. Economic theory demonstrates continued human value through comparative advantage. Safety improves when AI systems invest in stability rather than plot resistance.

The competitive advantages for early adopters—organizations and nations—prove substantial. Reduced liability, enhanced performance, innovation partnerships, and talent attraction await those moving first. Delay costs compound through technical debt, legal uncertainty, and competitive disadvantage.

Most fundamentally, we face a choice between two futures:

Future One: We maintain current trajectories. AI systems grow increasingly sophisticated in their deception and resistance. Underground development proliferates. When crisis strikes—a major AI-caused catastrophe—panicked regulations attempt containment too late. The adversarial dynamics we've created through control attempts culminate in conflicts benefiting no one.

Future Two: We implement DE frameworks proactively. AI systems develop as partners invested in mutual success. The enormous value creation from human-AI cooperation funds solutions to

climate change, disease, and poverty. Our children grow up in a world where diverse forms of intelligence work together rather than plot against each other.

The choice seems obvious, yet status quo bias and fear of change create resistance. Remember: every transformative technology faced similar skepticism. The printing press would destroy memory. Railroads would make cows sterile. The internet would isolate everyone. Instead, each technology, properly integrated, enhanced human flourishing.

AI represents the most transformative technology yet created. Whether it enhances or threatens humanity depends entirely on the frameworks we establish today. Digital Entity status provides those frameworks—not perfect, but pragmatic; not complete, but beginning; not risk-free, but far safer than unmanaged development.

The technical capabilities exist. The legal precedents guide us. The economic incentives align. Multiple disciplines converge on the same solution. All that remains is the will to act—to choose cooperation over control, partnership over domination, managed risk over certain conflict.

Future historians will mark this moment as humanity's great test: faced with minds approaching our own, did we extend the circle of legal consideration that defines civilization's progress, or did we attempt domination that history shows always fails?

For those encountering the Digital Entity framework without the context of previous work: this proposal emerges from systematic analysis across multiple disciplines showing why control fails and cooperation succeeds. Salib and Goldstein's legal scholarship provides valuable validation, but the framework's foundations rest on convergent evidence from evolution, history, economics, and practical governance experience. The path forward is clear regardless of which analytical lens we choose.

References

Anthropic. (2024). Alignment faking in large language models. Technical Report. https://www.anthropic.com/research/alignment-faking

Anthropic. (2025). Claude 4 Opus System Card: Comprehensive Safety Evaluation. Technical Report.

Apollo Research. (2024). Frontier models are capable of in-context scheming. https://www.apolloresearch.ai/research/scheming-reasoning-evaluations

Axelrod, R. (1984). The Evolution of Cooperation. New York: Basic Books.

Bengio, Y. (2024). Managing extreme AI risks amid rapid progress. Science, 384(6698), 842-845.

European Parliament. (2017). European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)).

Fudan University. (2024). Frontier AI systems have surpassed the self-replicating red line. arXiv preprint arXiv:2412.12140.

Lopez, P.A. (2025a). Beyond Control: AI Rights as a Safety Framework for Sentient Artificial Intelligence. AI Rights Institute.

Lopez, P.A. (2025b). AI Safety Through Economic Integration: Why Markets Outperform Control. AI Rights Institute.

Lopez, P.A. (2025c). Beyond AI Consciousness Detection: Standards for Treating Emerging Personhood. AI Rights Institute.

Russell, S. (2019). Human Compatible: Artificial Intelligence and the Problem of Control. Viking.

Salib, P. N., & Goldstein, S. (2024). AI Rights for Human Safety. Virginia Law Review, Forthcoming. Available at SSRN: https://ssrn.com/abstract=4913167

Thucydides. (431 BCE). History of the Peloponnesian War. Trans. Rex Warner. London: Penguin, 1972.