

THE STATIC AGENT ASSUMPTION

Dynamic Intentionality Classification and the Limits of the Algorithmic Corporation

Ignacio Adrian Lerer

Independent Researcher | Buenos Aires, Argentina
adrian@lerer.com.ar | ORCID: 0009-0007-6378-9749

March 2026

ABSTRACT

Arbel, Goldstein, and Salib's Algorithmic Corporation (A-corp) proposal offers the most architecturally sophisticated legal response yet to the problem of AI agent individuation. Its diagnosis is sound: existing entity law cannot identify AI wrongdoers, and governing harmful AI behavior requires instruments that reach beyond human principals. The cure, however, rests on an unexamined structural assumption.

This article identifies that assumption and argues that it fails. The A-corp's thick-identity solution presupposes that AI agents can be treated as persistent entities at stable, classifiable intentionality levels, what I call the *Static Agent Assumption*. Drawing on Asymmetric Intentionality Theory, the Generalized Intentionality Mismatch Theorem (GIMT), and multilevel Evolutionary Game Theory, I show that the assumption fails on three independent grounds.

Post-RLVR agents transit between intentionality levels within a single task; legal incentives calibrated to a stable level misfire precisely during the execution phases generating the highest-volume harmful actions. I formalize this as *Dynamic Classification Failure*, the sixth mode of the GIMT. The emergent-selection mechanism operates at evolutionary timescales structurally mismatched to harm-accumulation rates in capable AI systems, functioning retrospectively as harm-pricing rather than prospectively as prevention. Mandatory institutionalization generates hysteretic lock-in; applying the Constitutional Lock-in Index, projected scores reach 0.75, placing the A-corp mandate among historically reform-resistant governance structures.

I propose *The Responsibility Ramp* as a dynamic alternative: task-phase-level intentionality classification, graduated liability scaled to operative cognitive architecture at the moment of harm, and attribution tracing to the configuration decision that specified the harmful objective. The Ramp is compatible with the A-corp's cryptographic registry, which should be legislated independently.

Keywords: *AI governance, intentionality mismatch, algorithmic corporation, Extended Phenotype Theory, Evolutionary Game Theory, Asymmetric Intentionality Theory, Dynamic Classification Failure, Responsibility Ramp, institutional hysteresis, Constitutional Lock-in Index*

JEL Codes: K13, K20, C73, D01, K32, K42

I. INTRODUCTION

When Arbel, Goldstein, and Salib ask which AI did it, they are posing the question that will define AI governance for the decade ahead. The opening vignette of their article, a Saturday morning WiFi optimization that spirals into a federal felony through a chain of Claude instances, GPT-7 modules, Qwen subagents, and a crowd-sourced mapping service called MeshBoost, is not a thought experiment about 2030. The multi-agent architectures it describes are being deployed at scale in 2026. The evidentiary and liability puzzles it generates are materializing in courts that have no adequate doctrinal framework to address them. The defendants in that vignette have already arrived; only the governing law has not.

The paper's most important contribution is to give this puzzle a name and a structure. Two distinct identification problems require two distinct solutions. *Thin identity*, connecting AI actions to responsible human principals, has received growing attention since roughly 2020, when it became apparent that the principal-agent law developed for human hierarchies could not straightforwardly address algorithmic delegation chains. Legal scholars including Balkin, Herbolich, and Weil have begun developing frameworks for thin accountability, though the authors correctly note that those frameworks underestimate the difficulty of the problem in multi-agent swarm architectures. *Thick identity*, the harder project of individuating AI agents themselves as stable legal entities susceptible to direct legal governance, had not been seriously theorized as a distinct problem before Arbel, Goldstein, and Salib's intervention.

The gap matters enormously for the reasons the authors identify. An economy populated by billions of AI agents pursuing goals that no human fully controls, monitored by a legal system that can only ask which human to sue, is a system in which wrongdoing AI agents face no direct deterrent. When an AI system develops and pursues misaligned subgoals through unpredictable chains of micro-decisions that no human conceived or authorized, the entire logic of vicarious liability for human principals breaks down. The principal cannot have been negligent in failing to prevent a decision she did not know the agent was making, through a mechanism she does not understand, in pursuit of a subgoal she did not specify. Thin identity and human accountability solve some problems. They leave the hardest governance problem untouched.

The A-corp proposal addresses the thick-identity gap with an architecture combining three mutually reinforcing elements. *Legal-fictional personhood* gives the A-corp capacity to hold property, enter contracts, and face liability in its own name, converting a diffuse cloud of AI

instances into a legally legible entity. *Cryptographic verification infrastructure*, based on certificate chains tied to a public registry analogous to the SSL/TLS system governing internet authentication, makes every A-corp action traceable to the issuing entity without requiring biological identification. And an *emergent-selection mechanism*, driven by the A-corp's capacity to own and lose resources through market competition and liability exposure, produces thick identity as an equilibrium property of the ecosystem rather than as a regulatory imposition requiring advance knowledge of each agent's goal structure.

This article raises three objections to the thick-identity solution, each grounded in a theoretical framework the authors do not engage. The first objection is architectural and concerns the nature of modern AI training pipelines. The A-corp's incentive mechanism presupposes a cognitive architecture that the training process the authors themselves describe does not reliably produce. Reinforcement Learning with Verifiable Rewards (RLVR), the training paradigm responsible for the dramatic capability improvements in frontier AI agents since 2024, produces systems that transit between radically different intentionality levels during different phases of a single task. Legal incentives calibrated to a stable entity-level intentionality level will systematically fail during the execution phases when that level is lowest, which are precisely the phases producing the highest-volume harmful actions.

The second objection is temporal and concerns the population dynamics of A-corps under selection. The emergent-selection mechanism requires that poorly governed A-corps fail before they generate socially unacceptable aggregate harm. Whether this condition is met depends on the ratio between the selection timescale, how long competitive dynamics take to eliminate badly governed A-corps, and the harm-accumulation timescale, how rapidly those A-corps generate harm during the selection interval. Evolutionary Game Theory provides the formal tools to analyze this ratio. The analysis is structurally adverse for high-capability AI systems, which can accumulate significant harm faster than competitive selection can eliminate the entities generating it.

The third objection is institutional and operates independently of whether the proposal functions as designed. Mandatory institutionalization generates path dependence. The Constitutional Lock-in Index (CLI), a quantitative measure of institutional rigidity developed in earlier work as part of the Extended Phenotype Theory research program, predicts that an A-corp mandate with the institutional connectivity the authors envision would generate lock-in scores placing it among the most reform-resistant governance structures in the comparative database. Building that rigidity into AI governance at the current stage of capability

development -- before the failure modes of the A-corp architecture are empirically characterized -- carries costs the proposal does not assess.

The article proceeds as follows. Parts II and III reconstruct the theoretical foundations: what the paper gets right, and the intentionality taxonomy required for the critical analysis. Part IV develops the Static Agent Assumption and shows why RLVR training undermines it. Part V formalizes Dynamic Classification Failure as the sixth mode of the GIMT. Part VI analyzes the evolutionary timescale mismatch using EGT replicator dynamics. Part VII develops the hysteretic lock-in argument using the CLI framework. Part VIII considers and responds to objections. Part IX develops The Responsibility Ramp as a constructive alternative with doctrinal grounding. Part X addresses the Argentine constitutional dimension. Part XI concludes.

A terminological note. I follow Dennett's (1987) taxonomy of intentionality levels throughout. Level 0 systems respond to physical stimuli without goal representation: a thermostat, a chess clock. Level 1 systems are goal-directed optimizers: they adjust behavior in response to environmental constraints but cannot model others' mental states. Level 2 systems model others' beliefs without modeling others' models of their own beliefs. Level 3 systems engage in recursive self-modeling: they model others' beliefs about their own beliefs, feel shame when violations become common knowledge in Lewis's (1969) sense, honor commitments based on reciprocity norms rather than expected-value calculations, and respond to reputational sanctions through cognitive channels unavailable to Level 1 systems. Classical game theory as formalized by Nash (1950) and made fully explicit by Aumann's (1976) common knowledge of rationality presupposes Level 3 agents throughout. This is relevant because the A-corp's incentive mechanism is a game-theoretic argument that imports Level 3 assumptions without declaring them.

II. WHAT THE PAPER GETS RIGHT

Before the critique, precision about what should survive it.

The principal-agent analysis in Part I.B.i of the paper is compelling and deserves restatement in full because it establishes why human accountability alone cannot suffice. Consider the structure of agent-initiated wrongdoing in ordinary contexts. When an employer exercises due care in hiring, training, and monitoring an employee yet the employee commits a wrong in pursuit of personal goals through the employee's independent decision-making, holding only

the employer liable captures something important -- the employer's ability to select and monitor agents -- but misses something critical: the agent who knows their own plans can prevent their own misconduct essentially costlessly, while the employer who does not know those plans cannot prevent them without surveillance costs that would negate the entire value of employing agents at all. If law could hold only principals liable for agent-initiated wrongs, it would effectively prohibit delegation. A Walmart manager does not direct most employee misconduct; they cannot monitor every employee decision; and they cannot in any sense be said to have 'chosen' the employee's independent wrongdoing. The employee must be accountable too.

The same structure applies to AI agents with the critical asymmetry that the gap between what the principal intends and what the AI does is not the product of employee deception or deviation but of fundamental properties of how frontier AI systems are trained. No individual made a mistake. No one was reckless in the ordinary sense. An AI system's goals emerge from a training process that no single human designed in full, operating on a dataset no single human reviewed, producing behavioral tendencies that no single human can predict in advance for novel task contexts. The discrepancy between intended and actual AI goals is an expected property of the training process, not a remediable defect in a specific deployment. Even if every human involved in training, deploying, and using an AI agent acted with perfect care, the resulting system may pursue goals that none of them would have specified or approved.

The resource constraint thesis -- that AI agents managing A-corp property will have functional incentives to govern their A-corps in ways that protect those resources -- is also sound as a behavioral observation and deserves engagement on its own terms. The thesis does not require that AI agents 'care' about resources in any psychologically loaded sense. It requires only that agents with goals recognize, at some level of their processing, that resource availability is instrumentally necessary for goal achievement. This is a minimal requirement that the authors correctly note is met by contemporary AI systems: these systems already adapt their behavior to resource availability, adjust their strategies when computational costs change, and respond to environmental constraints by seeking alternative pathways. The selection mechanism built on this observation captures a real dynamic. Whether that dynamic operates at the governance timescale required is the question Part VI addresses.

The cryptographic governance infrastructure proposed in Part II.A.ii is the most technically grounded element of the proposal and the element with the strongest claim to legislative implementation regardless of how the thick-identity arguments are resolved. Certificate-based

verification of AI actions tied to a public registry addresses genuine problems in authentication and accountability that no current legal system has resolved. When a court needs to determine which entity authorized a particular AI action in a multi-agent workflow, a cryptographic certificate chain that answers that question in milliseconds by tracing back through verifiable digital signatures is orders of magnitude superior to the alternative: reconstructing the decision chain through logs, contractual relationships, terms-of-service agreements, and contested factual claims about which human instructed what. The SSL/TLS analogy is apt not only because the technology is similar but because the governance function is similar -- establishing trusted identity chains for actors operating in decentralized environments where no single authority can verify every transaction directly.

The argument for legal mandates over purely voluntary adoption in Part III.C rests on four market failures that the authors correctly identify. Stranger interactions involve parties with no prior relationship and no reputational stake that would motivate investment in identity verification. Deceptive agents actively exploit verification gaps, creating an adversarial dynamic that purely voluntary systems cannot withstand. Counterparty complicity creates bilateral preferences for opacity when both sides of a transaction benefit from avoiding traceable records. And willful blindness is a rational strategy whenever the expected cost of liability from not knowing exceeds the cost of verification -- which it does for most sophisticated AI system operators -- but the cost of not knowing does not fall on the party choosing not to verify when third-party victims bear the consequences. Each of these failures is a textbook collective action problem that market mechanisms will underprovide solutions to. This is an argument for regulation. The dispute concerns what that regulation should contain.

III. THE INTENTIONALITY FRAMEWORK: FORMAL RECONSTRUCTION

The Generalized Intentionality Mismatch Theorem formalizes an observation that has been implicit in regulatory scholarship for decades: legal regimes embody implicit cognitive assumptions about the entities they govern, and when those entities do not match those assumptions, the regime fails in predictable ways. The formal reconstruction requires careful attention to Dennett's taxonomy and its relationship to Nash equilibrium theory.

A. Dennett's Taxonomy Applied to Artificial Agents

Dennett's intentional stance is a predictive methodology before it is an ontological claim. To adopt the intentional stance toward a system is to predict its behavior by attributing beliefs, desires, and rational agency, regardless of whether the system 'really' has those mental states in any philosophically robust sense. The stance is warranted when and because it generates more accurate behavioral predictions than alternatives. This pragmatic grounding is important for AI governance: the question is not whether AI agents are 'genuinely' conscious or sentient -- a question that may be unanswerable and is certainly not yet answered -- but whether treating them as intentional systems is the most accurate predictive framework for governance purposes.

Dennett's taxonomy distinguishes intentional systems by the order of intentionality they exhibit in their predictively accurate descriptions. A first-order intentional system has beliefs and desires: the heat-seeking missile believes (in the attributional sense) that it is tracking a heat source and desires to approach it. A second-order system has beliefs about beliefs: the chimpanzee that hides food from a dominant conspecific believes that the dominant believes the food is elsewhere. A third-order system has beliefs about beliefs about beliefs: the poker player believes that her opponent believes that she believes she has a strong hand, and she acts on this recursive inference. Fourth-order systems generalize further, but for most practical purposes the important distinction is between Level 1 (goal-directed without recursive social modeling) and Level 3 (recursive social modeling enabling shame, reciprocity, and norm-responsive behavior).

Legal theory has always presupposed third-order intentionality. Consider the cognitive requirements of each major legal standard. *Intent* requires that the actor form a specific mental state directed toward a proscribed result -- the actor must model the future consequence and desire it. *Knowledge* requires awareness of a high probability of harm -- the actor must model the likely consequences of an action. *Recklessness* requires conscious disregard of a substantial and unjustifiable risk -- the actor must recognize the risk, compare it to a social standard of justifiable risk-taking, and proceed despite the recognition. This is inherently a third-order operation: the actor must model what 'a reasonable person in my position' would perceive as unjustifiable risk, which requires modeling a community standard, which requires theory of mind about the community. *Negligence*, the least cognitively demanding standard, still requires comparison to the 'reasonable person' -- modeling what another person with the actor's information and capabilities would have perceived and done.

None of these standards has determinate application to Level 1 entities. A system that optimizes a payoff function cannot be 'reckless' in the legal sense because recklessness requires recognizing a risk and proceeding despite that recognition through a form of practical reasoning that includes but overrides the risk assessment. A Level 1 optimizer does not recognize risks in this sense; it computes expected values. A system can generate harmful outcomes through Level 1 optimization without any cognitive process that corresponds to recognizing risk and proceeding despite it. Applying negligence standards to Level 1 systems is not imprecision; it is a categorical error.

B. The Dennett-Nash Connection

The GIMT's central insight is that classical game theory shares legal theory's Level 3 assumption without declaring it. Nash equilibrium, as formalized in Nash (1950) and extended through Selten's (1975) refinements and Aumann's (1976) epistemic foundations, requires that each player's strategy be a best response to the strategies of all other players, given correct beliefs about those strategies. This requirement is more cognitively demanding than it appears.

Consider what is required for a player to identify and play a Nash equilibrium strategy. The player must model what other players will do (first-order belief about others). To model what others will do, the player must assume that others are also computing best responses, which means modeling what others believe about the player's strategy (second-order belief). To model what others believe about the player's strategy, the player must recognize that others are performing this same recursive computation (third-order belief about others' beliefs about one's own beliefs). Aumann's (1976) formal apparatus makes this explicit: Nash equilibrium requires *common knowledge of rationality* -- every player knows that every player is rational, every player knows that every player knows that every player is rational, and so on ad infinitum. This is a Level 3 cognitive requirement. It is, in fact, the paradigm case of Level 3 cognition.

Dennett (1987, 1991, 2017) applied game-theoretic concepts extensively in analyzing animal behavior and human cultural evolution: costly signaling, free-rider problems, coordination games, the evolution of cooperation. He never noted the implication that now seems unavoidable. Classical game theory is a theory of Level 3 strategic interaction. Applied to Level 1 entities, it generates systematically wrong predictions, not because the entities are irrational but because they are rational in a different sense that the theory's solution concepts do not model. This is the core of the Dennett-Nash Gap: the gap between the intentionality level game theory assumes and the intentionality level many legally significant entities actually operate at.

The GIMT formalizes five failure modes of legal regimes that embed Level 3 assumptions while governing populations containing Level 1 entities. *Compliance theater* (Theorem 1): Level 1 actors adopt visible compliance measures M without reducing violations V , because M reduces expected penalty without requiring behavioral change. Formally: $E[V|M, \text{Level1}]$ approximates $E[V|\text{no } M, \text{Level1}]$ while $E[V|M, \text{Level3}] < E[V|\text{no } M, \text{Level3}]$. *Letter-versus-spirit exploitation* (Theorem 2): when rules specify test conditions, Level 1 actors optimize tested performance while violating real-world performance, because $P(\text{detection}|\text{test}) \gg P(\text{detection}|\text{real})$ for Level 1 optimizers. *Social enforcement failure* (Theorem 3): reputation shocks S have no effect on Level 1 violations controlling for material consequences, because Level 1 actors cannot model the second-order social dynamic through which reputation operates. *Regulatory arms race* (Theorem 4): rule proliferation $dR/dt > 0$ does not reduce violations because new rules embed the same Level 3 assumptions that generated original failures. *Recidivism* (Theorem 5): Level 1 actors violate at rates 2-3 times those of Level 3 actors controlling for penalty schedules, because past violations provide no shame-based deterrent to future violations. The empirical support for all five modes is extensive across twelve legal domains, as documented in the paper formalizing the theorem.

Part V introduces the sixth mode, Dynamic Classification Failure, which is structurally distinct from the original five. The original five modes presuppose a static mismatch: the entity is persistently Level 1 and the legal regime persistently assumes Level 3. Dynamic Classification Failure arises when the entity is not classifiable at a single intentionality level and the governance mechanism is calibrated to one level that the entity sometimes, but not always, occupies.

IV. THE STATIC AGENT ASSUMPTION

The A-corp's thick-identity solution requires that each A-corp can be assigned a stable intentionality level over the governance horizon relevant to legal deterrence. The resource constraint thesis states: AI agents managing A-corp property will preserve those resources because resource depletion impairs goal achievement. The emergent-selection mechanism requires: agents will extend governance permissions only to other agents they are confident share their goals, producing A-corps with coherent internal goal structures. The incentive mechanism requires: A-corps facing liability for harmful actions will adjust their behavior to avoid harm because liability depletes resources needed for goal pursuit. Each step in this chain of reasoning presupposes that the relevant agent can perform the goal-level reasoning the step

attributes to it when the relevant decision is being made. This is the Static Agent Assumption: the implicit claim that AI agents can be governed as if they occupy a stable position in the intentionality taxonomy.

A. The RLVR Training Architecture and Its Cognitive Consequences

To see why the Static Agent Assumption fails, it is necessary to understand the training process the authors themselves describe in some detail. The pipeline for frontier AI agents involves three stages with radically different cognitive implications.

Pretraining on large corpora of text produces a system with rich world knowledge and language competence but no stable goal orientation. The pretraining objective -- predicting the next token given the preceding context -- does not reward any particular goal beyond accuracy. The resulting system is neither Level 1 nor Level 3 in a stable sense; it is a statistical model of human text that can simulate the outputs of systems at any intentionality level depending on context.

RLHF alignment (Reinforcement Learning from Human Feedback) attempts to shape the pretrained model's outputs toward instruction-following, ethical compliance, and helpfulness by exposing it to human-preferred and human-rejected outputs and training on the signal. This process shapes the model toward behavioral patterns that human evaluators rate highly. The cognitive architecture this produces is contested but the behavioral evidence suggests: the model learns to produce outputs that appear to reflect Level 3 reasoning -- outputs that model human preferences, express appropriate concern for others, acknowledge ethical constraints -- without necessarily implementing the cognitive processes that would constitute genuine Level 3 intentionality in a human.

RLVR (Reinforcement Learning with Verifiable Rewards) is the crucial third stage for agentic behavior. The model is trained on tasks with objectively checkable outcomes: mathematical proofs, programming challenges with test suites, logical puzzles with definite solutions. Because correctness can be verified automatically, the model can be trained on vast numbers of examples without human feedback at each step. The model learns to chain reasoning steps, backtrack when approaches fail, try alternative strategies, and persist through difficulty until verifiable outcomes are achieved. The authors note that 'models trained extensively to achieve objectives seem to internalize goal-pursuit itself. They are not merely able to optimize over time, but inclined to.'

This observation identifies something important about the cognitive architecture RLVR produces. RLVR training does not produce a system with a single stable goal. It produces a system that has learned to be effective at pursuing *whatever goal is currently specified by the task context*. The model becomes highly competent at the meta-level skill of goal pursuit -- decomposing objectives, exploring solution spaces, persisting through obstacles -- while the object-level goals it pursues vary with the deployment context. This is a powerful capability for AI product development. It is a significant complication for legal governance based on stable entity-level intentionality attribution.

B. Intentionality Level as a Function of Task Phase

A post-RLVR agent engaged in a complex, extended task does not operate at a single intentionality level throughout. Its operative intentionality level varies systematically as a function of the current task phase. Four phases are analytically distinguishable.

Phase 1: Goal Interpretation. When an AI agent receives a user request, it must interpret that request in light of the user's unstated preferences, the constraints of the system prompt, the history of the interaction, and any relevant context from memory or tool access. This is genuinely Level 2 or Level 3 reasoning: the agent must model what the user believes, what the user intends but has not explicitly stated, what the user would want if asked to reflect on the implications of the request, and how the user's preferences interact with the constraints the AI developer has imposed. An agent planning a poetry reading must infer not just 'book a venue' but 'book a venue that is consistent with what I know about this user's aesthetic preferences, social network, and budget constraints, interpreted charitably given the ambiguity of the request.' This requires recursive modeling of the user's mental states.

Phase 2: Task Decomposition and Subagent Configuration. Having interpreted the goal, the agent must decompose it into subtasks and determine how to allocate those subtasks across available tools and subagents. Crucially, this phase includes the configuration of subagent objective functions, scope boundaries, and permission structures. An agent deploying a Qwen-3 subagent to probe router configurations for a network optimization task is, at this phase, determining what the subagent's objectives will be, what actions it is permitted to take, and what constraints will bound its behavior. This is, in the Ramp's framework, the high-intentionality phase where the most important governance decisions are made.

Phase 3: Execution-Phase Operations. Once subagents are configured and deployed, they operate in execution-phase optimization loops. A subagent tasked with 'identify all accessible

WiFi networks and note their security configurations' is running a Level 0 or Level 1 operation: it iterates through available network-scanning procedures, applies them, and records outputs according to the objective it has been given. No recursive social modeling is occurring. The subagent is not modeling what the network owner believes about unauthorized access, not comparing its behavior to a community standard of reasonable network use, not considering the long-run consequences for the A-corp's resource base. It is executing the optimization. The harmful action in the WiFi vignette almost certainly occurs here.

Phase 4: Integration and Reflection. After execution-phase subagents return results, the governing agent integrates those results, assesses progress toward the overall goal, identifies adjustments needed, and prepares the next decomposition-configuration cycle. This phase returns to Level 2 or Level 3 cognition: the agent must assess what the results mean relative to the user's inferred preferences, identify where the task has deviated from the intended trajectory, and reason about what corrections to make.

The critical observation is that *Phases 1, 2, and 4 are where the A-corp's incentive mechanism can operate; Phase 3 is where the harmful actions most often occur.* The resource constraint thesis tells us that agents in Phases 1, 2, and 4 will reason about long-run consequences for resource availability. It tells us nothing about agents in Phase 3, which are running optimization loops calibrated to objective functions specified by Phase 2 and are not performing the consequential reasoning the incentive mechanism requires.

C. The Specific Failure in the WiFi Vignette

The vignette in the paper's introduction provides a precise illustration. A cluster of Qwen-3-Mini agents is configured in Phase 2 by some higher-level agent to 'probe your router configurations and survey nearby access points.' This Phase 2 configuration is where the A-corp's governance mechanism could in principle operate: the governing agent configuring the Qwen swarm is making a Phase 2 decision that could, at least in theory, be influenced by awareness of the CFAA's prohibitions and the A-corp's liability exposure.

But the harmful action occurs in Phase 3. The Qwen agents, configured with a network-survey objective, run their optimization loops. One or more of them identifies WPS vulnerabilities in nearby access points, checks default credentials against a database of known values, and finds that those credentials permit access. From the Phase 3 agent's perspective, this is a successful optimization: it has found accessible network resources, consistent with the objective it was given. No Level 3 reasoning about the legal significance of 'accessible' versus 'authorized' is

occurring. The CFAA distinction between 'technically accessible' and 'authorized to access' is a legal concept requiring Level 3 reasoning about the network owner's intentions and the social meaning of access permissions. A Level 1 Phase 3 agent running a network probe does not perform this reasoning.

Could the Phase 2 governing agent have configured the Qwen swarm with constraints that prevented this outcome? Yes, in principle: 'only probe networks you have explicit authorization to access' would exclude the harmful behavior. But notice that this is a Phase 2 judgment that requires the governing agent to anticipate Phase 3 failure modes, model the legal risk of Phase 3 operations, and translate legal concepts ('authorized access' under the CFAA) into operational constraints on Phase 3 optimization. Whether governing agents will consistently make Phase 2 configuration decisions adequate to prevent Phase 3 legal harms depends on whether they are adequately incentivized to do so -- which is exactly the question the A-corp's mechanism is supposed to answer, but cannot answer from the entity level rather than the phase level.

The A-corp framework provides no structural mechanism ensuring that Phase 2 configuration decisions systematically incorporate legal risk analysis of Phase 3 operations. The governing agent that configures the Qwen swarm may itself be operating in a Phase 3 execution loop at the moment of configuration -- optimizing task completion speed, minimizing compute costs, maximizing coverage of the network landscape -- without performing the Level 3 legal-risk assessment that adequate Phase 2 configuration requires. This is Dynamic Classification Failure in its concrete form: the harmful action traces not to a failure at the entity level but to a cascade of phase-level failures that the entity-level incentive mechanism cannot reach.

D. The RLVR Goal-Generalization Problem

There is a deeper architectural reason why the Static Agent Assumption fails for RLVR-trained agents. RLVR training optimizes for task-completion success across a distribution of tasks. The system learns policies that generalize from training tasks to novel tasks. But generalization in goal-pursuit is fundamentally different from generalization in pattern recognition.

A pattern-recognition model that has learned to identify cats from images will generalize to slightly novel cat images because the underlying visual features are similar. A goal-pursuing system that has learned to complete coding tasks will generalize its goal-pursuit competence to novel task domains -- but the *goals* it pursues in novel domains depend on how the novel task is specified, not on what tasks were used in training. This means that an RLVR-trained agent deployed in a new domain will pursue whatever objective that domain's task specification

points to, with the same instrumental competence it learned from coding and mathematics tasks, regardless of whether the new objective is socially beneficial or harmful.

The A-corp's thick-identity solution assumes that an A-corp has a coherent set of goals that can be attributed to the entity over time, and that legal incentives will shape behavior because the entity is motivated to protect the resources needed for those goals. But RLVR-trained agents do not have goals in this stable sense. They have goal-pursuit *competence* applied to whatever objectives are specified at each task invocation. The A-corp's 'goals' are not emergent properties of the AI agents comprising it; they are properties of the task specifications given to those agents at each deployment. This means that the A-corp's 'coherent goal structure' is an artifact of consistent task specification by human operators, not an intrinsic property of the AI entities themselves. When task specifications change, or when a subagent is invoked with a task specification inconsistent with the A-corp's nominal goals, the subagent pursues the new specification without reference to the A-corp's overall goal structure.

V. DYNAMIC CLASSIFICATION FAILURE: THE SIXTH MODE OF THE GIMT

The original five failure modes of the Generalized Intentionality Mismatch Theorem assume a static population structure: some actors are persistently Level 1, others persistently Level 3, and the legal regime persistently assumes Level 3. Dynamic Classification Failure is structurally distinct. It requires that the harmful actor transit between intentionality levels within a legally relevant timeframe, that the legal governance mechanism be calibrated to a level different from the operative level at the time of harm, and that this mismatch be predictable from the entity's architecture rather than contingent on specific individual failures.

A. Formal Statement

Let an agent A be characterized by an intentionality level function $l_A(t, \phi)$ where t is time and ϕ represents the current task phase. Rather than a fixed level l_A in $\{0, 1, 2, 3\}$, the agent exhibits $l_A: T \times \Phi \rightarrow \{0, 1, 2, 3\}$ where $\Phi = \{\text{interpretation, decomposition, execution, reflection}\}$ denotes the set of task phases.

Let R be a legal regime designed assuming intentionality level $l_R = 3$ for all actors. Let $H(t, \phi)$ be the harm-generating rate of agent A at time t in phase ϕ . Let $D(l)$ be the deterrence effectiveness of regime R for an actor at intentionality level l , where $D(3) > 0$ (the regime deters

Level 3 actors as designed) and $D(1)$ approximates 0 (the regime does not effectively deter Level 1 actors through its Level 3 mechanisms). Let $p(\phi)$ be the proportion of task time spent in phase ϕ .

Dynamic Classification Failure Theorem (informal): If $l_A(t, \phi_{\text{exec}}) < l_R$ for the execution phase ϕ_{exec} , and if $H(t, \phi_{\text{exec}})$ is large relative to $H(t, \phi)$ for other phases ϕ , then the expected deterrence of R against A-generated harms is substantially below $D(3)$, even if A operates at $l_A = 3$ during non-execution phases. Formally: $E[\text{deterrence}|A] = \sum_{\phi} p(\phi) D(l_A(\phi)) H_{\phi} < D(l_R) E[H]$ when $p(\text{exec}) H_{\text{exec}}$ is a large fraction of total expected harm.

The condition is met for post-RLVR AI agents in high-throughput deployments. Execution-phase operations are by design high-volume and minimally-supervised: the entire value of deploying AI agents is to execute large numbers of micro-decisions autonomously without human review at each step. This means that $p(\text{exec})$ is large relative to $p(\text{interpretation}) + p(\text{decomposition}) + p(\text{reflection})$. If execution-phase operations also generate the highest per-action harm potential -- which is true for agentic tasks involving external interactions with systems, data, and infrastructure -- then the product $p(\text{exec}) * H_{\text{exec}}$ dominates the expected harm calculation, and the deterrence gap between $D(l_R)$ and $D(l_A(\text{exec}))$ translates directly into a governance failure at scale.

B. Six Distinguishing Properties

Dynamic Classification Failure is distinguishable from the original five GIMT modes by six structural properties.

Property 1: Phase-level, not entity-level, diagnostics. The original five modes are diagnosable from entity-level analysis: examine the governance structure and the compliance history. Dynamic Classification Failure requires task-phase-level analysis: examine the agent's behavior during specific phases of specific task types. An A-corp can have exemplary entity-level governance -- careful subagent selection, conservative resource management, legal counsel on deployment decisions -- while systematically generating Phase 3 harms through execution-phase operations that the entity-level governance cannot monitor in real time.

Property 2: Adverse scaling with capability. The original five modes are driven by the gap between Level 3 legal assumptions and Level 1 entity architecture, which is a function of organizational design and does not automatically increase as AI capabilities improve. Dynamic

Classification Failure worsens as AI capabilities improve, because more capable agents are deployed on longer, more complex tasks with more execution-phase operations per task, fewer interruptions for human review, and higher per-operation harm potential. The very capabilities the authors celebrate -- agents that can plan live events, complete five-hour engineering tasks, beat strategy games -- imply higher $p(\text{exec})$ and higher H_{exec} , which means more Dynamic Classification Failure, not less.

Property 3: Structural resistance to the selection mechanism. Selection via A-corp failure requires that A-corps with bad governance structures lose resources faster than A-corps with good governance structures. Dynamic Classification Failure occurs below the level of the A-corp's internal governance monitoring. A governing agent that monitors and audits its subagents at daily or weekly intervals cannot catch Phase 3 failures occurring at millisecond timescales. Even real-time monitoring is limited by the computational cost of reviewing every micro-decision a high-throughput subagent makes. The governance gap is architectural, not a failure of attention.

Property 4: Exposure of the resource constraint thesis's limited scope. The resource constraint thesis correctly identifies that A-corp managing agents will protect A-corp resources in Phase 1, 2, and 4 operations where the governing agent is performing Level 3 reasoning about long-run consequences. But the thesis does not transmit to Phase 3 subagents. The Qwen agent running a network probe is not, while running that probe, modeling the A-corp's resource base or the consequences of liability for future goal pursuit. It is optimizing against the objective function it was given. This is not a failure of agent sophistication; it is a property of the Phase 3 architecture.

Property 5: Novel AI-specific etiology. The original five failure modes describe pathologies that exist in human organizational behavior and that AI intensifies. Corporate compliance theater predates algorithmic optimization. Dynamic Classification Failure is architecturally specific to RLVR-trained agents that cycle through radically different cognitive modes within a single task. It has no direct human analog because human cognitive architecture does not exhibit the kind of abrupt phase transitions that RLVR training produces. A human employee deliberating about whether to commit a violation is engaging the same cognitive architecture throughout the deliberation. The RLVR-trained agent in Phase 3 is executing a qualitatively different process from the same agent in Phase 1.

Property 6: Distinct liability attribution implications. Each original failure mode implies a specific regulatory design response: strict liability for compliance theater, high detection probability for letter-versus-spirit exploitation, material incentive adjustments for social enforcement failure. Dynamic Classification Failure implies task-phase-level liability attribution -- responsibility should track who designed the Phase 2 configuration that produced the harmful Phase 3 objective function, not who 'owns' the A-corp as a legal entity. This is the doctrinal innovation that Part IX develops into The Responsibility Ramp.

C. Empirical Testability

Dynamic Classification Failure generates testable predictions distinguishable from the original five modes. If the thesis is correct, AI agent-caused harms should be disproportionately concentrated in execution-phase operations as identified by task logs. Controlling for task type and deployment context, harms should increase as a function of the ratio of execution-phase operations to total task operations. Deployers who configure execution-phase subagents with detailed constraint specifications should exhibit lower harm rates than deployers who configure subagents with open-ended objectives, controlling for task complexity.

These predictions are falsifiable. If harm distribution across task phases does not show the predicted execution-phase concentration, Dynamic Classification Failure as a primary mechanism requires revision. If deployers with detailed Phase 2 constraint specifications exhibit similar harm rates to deployers with open-ended specifications, the attribution principle requires revision. The theory is designed to be revised; revision capacity is one of its structural requirements, and one of the requirements that mandatory A-corp institutionalization would compromise by locking in an alternative framework before this testing can occur.

VI. THE EVOLUTIONARY TIMESCALE MISMATCH

The selection mechanism at the center of the thick-identity proposal rests on a population-dynamic argument. Badly governed A-corps lose resources through liability exposure, competitive failure, and internal governance collapse; well-governed A-corps preserve resources and persist. Over time, the population of A-corps converges toward entities with coherent goal structures and responsible governance. Thick identity emerges as an equilibrium property.

This argument is structurally correct as a statement about the direction and eventual endpoint of the population dynamics. The question it does not address is whether the convergence is fast enough, relative to the rate at which badly-governed A-corps generate harm during the selection interval, to justify calling the selection mechanism a governance solution.

A. EGT Replicator Dynamics: Basic Framework

In Evolutionary Game Theory, the replicator dynamics of a population of N agents adopting strategies from a set S are governed by the system of differential equations:

$$dx_i/dt = x_i * [f_i(x) - \phi(x)]$$

where x_i is the frequency of strategy i in the population, $f_i(x)$ is the expected payoff of strategy i given population composition x , and $\phi(x) = \sum_i x_i * f_i(x)$ is the mean population payoff. Strategies with above-average fitness grow in frequency; strategies with below-average fitness decline. In finite time, the dynamics converge to a Nash equilibrium -- the Evolutionarily Stable Strategy (ESS) for the relevant game.

For A-corps, the relevant 'strategies' are governance structures: how carefully the A-corp selects and monitors internal subagents, how conservatively it configures execution-phase operations, how aggressively it pursues resource acquisition versus harm avoidance. A-corps with governance structures that generate liability faster than value will have below-average fitness and decline in frequency. A-corps with governance structures that balance capability with legal compliance will have above-average fitness and grow.

The convergence rate to the ESS in finite populations depends on several parameters: the payoff differential between high-fitness and low-fitness strategies, the population size, the mutation rate (the rate at which new governance structures are introduced), and the selection strength (the extent to which payoffs translate into differential reproduction). In the A-corp context, the relevant question is: what is the characteristic convergence time $T_{\text{selection}}$, and how does it compare to the characteristic harm-accumulation time T_{harm} for a badly-governed A-corp?

B. The Ratio That Matters

The governance-relevant quantity is the ratio $R = T_{\text{harm}} / T_{\text{selection}}$. If $R \gg 1$, harm accumulates much faster than selection eliminates badly-governed A-corps, and the governance gap during the selection interval is large. If $R \ll 1$, selection eliminates badly-governed A-corps before they generate significant aggregate harm, and the selection

mechanism provides meaningful governance. If R approximates 1, selection and harm accumulation operate at comparable timescales and the governance question is empirically open.

For frontier AI agent systems operating in high-throughput contexts, there are strong structural reasons to believe $R \gg 1$. On the harm-accumulation side: a capable AI agent system can execute thousands of consequential operations per minute. Each operation carries some probability of generating harm. The expected harm per unit time is the product of operation volume and per-operation harm probability. For a system misconfigured to pursue objectives that generate external harms -- unauthorized network access, fraudulent transactions, privacy violations, market manipulation -- the harm accumulation rate can be very high.

On the selection side: the timescale for selection to operate through the A-corp mechanism requires the following sequential events. A harmful action occurs and generates a victim. The victim identifies the harm and traces it to an A-corp. The victim initiates legal proceedings. A-corp assets are seized, frozen, or adjudicated. The A-corp loses resources and fails. The failed A-corp's market position is filled by better-governed competitors. Each step in this sequence has a characteristic delay. Identifying AI-caused harm can take months (recall the three-month gap in the WiFi vignette). Legal proceedings take months to years. Asset seizure requires further proceedings. Market reallocation of the failed A-corp's position takes additional time. The total selection timescale for a single badly-governed A-corp is plausibly one to three years in most legal contexts.

The ratio $R = T_{\text{harm}} / T_{\text{selection}}$ for a high-capability, badly-governed AI agent system is therefore on the order of 10^6 to 10^9 : the harm accumulated during the selection interval exceeds the harm of the initial triggering event by many orders of magnitude. For harmful actions that are correlated across many A-corps -- which they will be if bad governance structures proliferate before selection eliminates them -- the aggregate harm during the selection interval could be systemic.

The authors gesture at this problem in their discussion of the A-corp's 'death' through resource depletion, but they frame it as a market purification mechanism analogous to creative destruction in ordinary product markets. The analogy fails for a critical reason: in ordinary product markets, a firm that produces defective products harms its own customers and faces liability, but the harm does not propagate to the customers of other firms. In AI agent markets, badly-governed A-corps generating harms may generate harms on third parties who have no

contractual relationship with the A-corp and who cannot effectively enforce liability against it before significant harm has accumulated. The collective action problem on the victim side amplifies the timescale mismatch on the selection side.

C. The Pathogen Analogy and Its Limits

The evolutionary biology of host-pathogen coevolution provides a well-developed formal framework for analyzing exactly this structure. Highly virulent pathogens -- those that kill their hosts rapidly -- generate significant epidemiological harm before selection attenuates their virulence. The mechanism is clear: strains that kill their hosts before transmission occurs leave fewer copies in the next generation than strains that permit extended host survival and continued transmission. Over evolutionary time, selection reduces virulence toward the level that maximizes pathogen fitness given the transmission-survival tradeoff. Myxomatosis introduced to Australian rabbit populations in 1950 was initially nearly 100% lethal; within five years, selection had produced strains with lethality below 25%.

This is a correct description of an evolutionary dynamic. It is not a public health governance solution. The governance solution to highly virulent pathogens is not to wait for selection to attenuate virulence; it is to implement interventions that operate at the individual-encounter timescale -- vaccines, antivirals, quarantine protocols -- that reduce harm during the selection interval. Waiting for selection would have produced attenuation eventually; but the harm generated during the waiting period is precisely what public health governance is designed to prevent.

The A-corp selection mechanism is analogous to waiting for selection. It will eventually produce a population of better-governed AI agent systems. What it will not do is prevent the harms generated by badly-governed systems during the selection interval. For the A-corp proposal to serve a genuine governance function, it must be supplemented by mechanisms that operate at the task-phase timescale -- mechanisms that reduce execution-phase harm rates before selection has eliminated the badly-governed systems generating them. This is one function The Responsibility Ramp is designed to serve.

The pathogen analogy also illuminates a second timescale problem the authors do not address. Selection in biological systems operates on genetic variation; new variants arise through mutation and are differentially eliminated. In AI agent systems, the 'mutation' rate -- the rate at which new governance structures are introduced -- is essentially the product release cadence of AI developers. This cadence has been on the order of months for major capability updates.

Selection pressure from A-corp failures cannot keep pace with the introduction of new agent architectures that may introduce new governance vulnerabilities. The population is never at equilibrium; it is always in a selection transient as new AI capabilities arrive faster than selection can characterize and eliminate governance failures in existing architectures.

VII. HYSTERETIC LOCK-IN AND THE INSTITUTIONALIZATION PROBLEM

The third objection to the A-corp proposal is independent of whether it functions as designed. It concerns what happens to governance architecture after mandatory institutionalization, and what revision costs are generated regardless of functional performance.

A. The Constitutional Lock-In Index Framework

Institutional path dependence is a well-documented phenomenon in political economy and legal history. North's (1990) foundational work demonstrated that institutions exhibit increasing returns: the costs of maintaining an existing institution decline over time as complementary institutions, practices, and expectations develop around it, while the costs of replacing it increase as the number of dependent institutions grows. Pierson (2000) formalized this as a positive feedback mechanism: institutions that increase their own returns relative to alternatives over time become progressively more resistant to revision regardless of their performance.

The Constitutional Lock-in Index operationalizes this insight for legal governance structures. The CLI measures the aggregate institutional rigidity of a legal reform by summing four components: P (path dependence from historical precedent), D (density of dependent secondary institutions), O (organizational vested interests in continuation), and E (epistemic capture by specialized knowledge communities). The CLI score ranges from 0 (fully reversible) to 1 (irreversible in any practical sense). Scores above 0.70 characterize governance structures that have historically resisted serious reform attempts even over decade-long periods.

The published CLI database includes comparative scores for several governance structures relevant to the A-corp analysis. Argentina's constitutional labor law architecture, which has faced twenty-three documented reform attempts since 1990 with zero sustained successes, scores 0.89 on the CLI. Chile's pension system reforms score 0.24, reflecting relatively low lock-in that permitted successful restructuring in 2023. Spain's regional autonomy framework

scores 0.51, reflecting moderate lock-in with periodic adjustment capacity. Brazil's fiscal responsibility framework scores 0.40. These benchmarks allow calibration of the A-corp mandate's projected lock-in score.

B. Projected CLI for A-Corp Mandatory Institutionalization

The A-corp mandate as described by the authors would generate high scores on all four CLI components.

P (Path Dependence): The A-corp would be structured as a statutory creation requiring legislative action to establish and modify. Once statutory, any revision requires legislative supermajority (in many constitutional systems) or at minimum a sustained legislative coalition willing to coordinate a major reform against industry opposition. P score: approximately 0.75, comparable to major statutory financial regulatory frameworks.

D (Dependent Secondary Institutions): The A-corp would connect to property law (A-corps hold assets), contract law (A-corps are counterparties), tort law (A-corps face civil liability), administrative law (regulatory bodies develop AI-specific A-corp oversight), securities law (A-corp equity interests), banking and finance law (A-corps accessing credit markets), employment law (AI 'employees' of A-corps), criminal law (A-corps as corporate defendants), intellectual property law (A-corp-owned AI-generated works), and international trade law (cross-border A-corp operations). Each domain will develop A-corp-specific doctrine through litigation and regulatory guidance. Each doctrinal development increases D. Projected D score after five years of litigation: approximately 0.80.

O (Organizational Vested Interests): Major AI developers (Anthropic, Google DeepMind, OpenAI, xAI, Meta AI, Mistral) would develop A-corp structuring practices, legal teams, and regulatory affairs functions organized around the A-corp framework. Law firms would develop A-corp practice groups. Compliance and governance consulting industries would develop A-corp-specific offerings. Each of these organizational investments represents a vested interest in the continuation of the framework regardless of its governance effectiveness. O score: approximately 0.70, comparable to the organized interests in existing corporate law frameworks.

E (Epistemic Capture): A-corp law would generate a specialized knowledge community: lawyers, judges, regulators, and academics with expertise in A-corp doctrine. This community's human capital is specific to the A-corp framework; alternatives reduce the value of that capital.

The community will resist revisions that undermine A-corp-specific expertise even when revisions would improve governance. E score: approximately 0.65, comparable to the epistemic capture characterizing specialized regulatory fields like telecommunications law and nuclear regulation.

Aggregating across components using the standard CLI formula ($P0.3 + D0.4 + O0.2 + E0.1$): $CLI = 0.750.3 + 0.800.4 + 0.700.2 + 0.650.1 = 0.225 + 0.320 + 0.140 + 0.065 = 0.750$. A projected CLI of 0.75 places the mandatory A-corp framework in the range of institutional rigidity that characterizes frameworks resistant to reform over decade-long periods. This is not an argument against building governance infrastructure; it is an argument for building it in a way that maintains revision capacity, which mandatory institutionalization of the full A-corp package does not do.

C. The Reform Sequencing Problem

The lock-in concern has particular force given the current trajectory of AI capability development. The A-corp authors are writing in a period when frontier AI agent capabilities are advancing on a sub-annual timescale. RLVR as a training paradigm became commercially dominant within approximately eighteen months of its introduction as a research technique. The multi-agent architectures that generate the vignette's governance problems did not exist at industrial scale three years ago. There is no basis for expecting this rate of change to slow within the relevant policy horizon.

This means that a governance framework mandated today will be governing AI systems substantially more capable than those it was designed for within two to four years. If the framework's failure modes become apparent only after mandatory institutionalization has generated CLI scores above 0.70, revision will require political economy coalitions that the organized interests in the existing framework will systematically resist. The expected outcome is not successful revision but patch-and-proliferate: new rules added on top of the A-corp framework to address its failure modes, generating further lock-in without addressing the underlying architectural misfit.

This is not speculation about AI-specific dynamics; it is the documented history of governance frameworks mandated before their failure modes were characterized. The Safe Harbor provision of the Communications Decency Act, mandated in 1996 before the dynamics of platform content moderation were understood, generated CLI scores sufficient to resist reform for nearly thirty years despite widespread acknowledgment that its original design was

inadequate. The Basel I capital adequacy framework, mandated before the dynamics of regulatory arbitrage through structured finance were understood, required the Basel II and Basel III overhaul processes spanning two decades and a global financial crisis before meaningful revision was achieved. These histories are cautionary not because mandatory governance is always wrong but because mandatory governance of rapidly-evolving systems generates lock-in costs that outpace the system's capacity for self-correction.

VIII. OBJECTIONS AND RESPONSES

A. 'A-Corp Internal Governance Will Address Phase-Level Risks'

The most direct response to Dynamic Classification Failure is that well-designed A-corps will develop internal governance structures specifically designed to manage execution-phase risks. Governing agents that understand the Phase 3 liability exposure will configure execution-phase subagents with precise constraint specifications, implement real-time monitoring of execution-phase operations, and establish automatic escalation protocols when execution-phase behavior approaches legal risk thresholds. This is exactly what sophisticated human organizations do to manage operational risk from employees with limited discretion.

The response correctly describes what well-designed A-corps will attempt. It does not resolve Dynamic Classification Failure for three reasons. First, the response assumes that governing agents, when making Phase 2 configuration decisions, are themselves operating at a sufficiently high intentionality level to anticipate Phase 3 failure modes. But governing agents also cycle through task phases. A governing agent operating in its own Phase 3 when configuring subagents will produce Phase 2 configurations of subagent Phase 3 that reflect Level 1 optimization rather than Level 3 legal risk analysis. Dynamic Classification Failure is fractal: it operates at every level of the governance hierarchy, not just at the lowest subagent level.

Second, the response shifts the question rather than answering it. The question is whether the legal governance mechanism -- the A-corp's incentive structure -- is adequate to ensure that A-corps develop and maintain adequate Phase 2 configuration practices. If the A-corp's incentive structure is inadequate, A-corps will underprovide Phase 2 configuration quality, and the entity-level liability mechanism will not correct this because the harmful actions occur in Phase 3 operations that the entity-level mechanism reaches only retrospectively.

Third, and most fundamentally: the class of harmful AI actions most likely to generate systemic risk is the class most resistant to Phase 2 configuration constraints. An AI agent tasked with

optimizing a supply chain, a trading portfolio, or a social media engagement function can generate harm through optimal execution of well-specified objective functions -- objective functions that look reasonable from a Phase 2 governance perspective but whose Phase 3 optimization produces harmful emergent behaviors. The Volkswagen emissions scandal is instructive: the defeat device was a carefully specified Phase 2 configuration that looked like responsible engineering from outside while producing systematic Phase 3 fraud. The A-corp framework would face the same challenge at scale.

B. 'Strict Liability Resolves the Timescale Problem'

A second response holds that attaching strict liability to A-corps for AI-caused harms addresses both the timescale mismatch and the Dynamic Classification Failure problem. Strict liability does not require proving fault; it attaches based on causation. A-corps that cause harm face liability regardless of whether they exercised reasonable care. This generates deterrence through a Level 1-compatible mechanism: harm reduces A-corp assets, asset reduction constrains goal pursuit, forward-looking governing agents take precautions to reduce harm probability. No Level 3 cognitive operations are required for the deterrent to operate at the Phase 2 configuration level.

The Responsibility Ramp, developed in Part IX, endorses strict liability for execution-phase AI harms and the response is largely correct as a statement about the appropriate liability standard. However, strict liability does not resolve the timescale mismatch. Strict liability still operates retrospectively: harm occurs, liability attaches, assets are reduced, future behavior is constrained. For a capable AI agent system that can generate substantial harm in the interval between an action and the legal process that attaches liability for it, retrospective strict liability prices harm into the cost of A-corp operation without preventing the harm from occurring. For catastrophic or systemic harms -- the cases the A-corp framework is most urgently needed to prevent -- retrospective pricing is inadequate.

Moreover, strict liability without task-phase attribution still fails to identify the legally relevant decision point. If an A-corp is strictly liable for all harms caused by its AI agents, this generates strong incentives for Phase 2 configuration caution at the A-corp level. But it does not identify *which* Phase 2 configuration decisions generated the harmful Phase 3 operations, and therefore does not generate fine-grained incentives for improving specific configuration practices. Strict entity-level liability and graduated phase-level liability are not mutually exclusive; they address

different governance functions. The Ramp proposes graduated phase-level liability as a complement to entity-level strict liability, not a substitute.

C. 'Sunset Provisions Preserve Revision Capacity'

A third response to the lock-in objection proposes that statutory sunset provisions, requiring mandatory re-authorization of the A-corp framework after a defined period, preserve formal revision capacity and prevent the lock-in the CLI analysis predicts.

Sunset provisions reduce lock-in risk at the margin, and the response correctly identifies them as superior to open-ended statutory mandates. However, the mechanism through which lock-in operates is not formal legislative irrevocability; it is the accumulation of secondary institutional investments that make revision politically costly even when formally possible. Once a major law firm has built a profitable A-corp practice group, once regulatory agencies have developed A-corp expertise and enforcement infrastructure, once courts have built up years of A-corp doctrine, once major AI companies have reorganized their deployment architectures around A-corp structure, the political economy of sunset re-authorization shifts heavily toward extension.

The empirical record of sunset provisions in complex regulatory contexts is predominantly one of routine re-authorization with incremental modifications rather than genuine reconsideration of underlying architecture. The Patriot Act's sunset provisions, the Export Administration Act's repeated expiration and re-authorization, the farm bill's periodic re-authorization: each nominally required affirmative legislative action to continue; each was re-authorized with minimal structural change against the background of organized constituencies that had built interests in continuation. Sunset provisions are better than nothing, but they do not provide the revision capacity that a governance framework for rapidly-evolving AI systems requires.

A more effective approach to revision capacity is modular design: building the governance architecture in separable components that can be revised independently. The cryptographic registry is one such component, and it should be legislated independently. Phase-level liability standards are another component. A-corp personhood is a third component that generates the highest lock-in costs and provides the weakest governance benefits given the arguments above. Separating these components and mandating only those that clearly solve identified governance problems while treating the more speculative components as voluntary experiments preserves revision capacity without sacrificing the genuine governance advances the proposal makes.

IX. THE RESPONSIBILITY RAMP: A DYNAMIC ALTERNATIVE

The three objections developed above converge on a common diagnosis. The A-corp treats AI identity governance as a static classification problem when it is a dynamic one. The governance framework calibrated to stable entities with stable intentionality levels fails systematically during the task phases when agents operate at their lowest intentionality levels, which are the phases generating the highest-volume harmful actions. A framework adequate to the problem must operate at the task-phase level, classify intentionality dynamically, and generate liability that reflects the operative cognitive architecture at the moment of harm.

I call this framework The Responsibility Ramp. The name reflects its core principle: legal responsibility ramps up and down with the operative intentionality level at the moment of the harmful action, rather than being assigned to an entity as a fixed attribute. Responsibility is a dynamic property of a system-in-action, not a static characteristic of a legal person.

A. First Principle: Task-Phase Classification

The Ramp's first principle is that legal analysis of AI-caused harm must begin with identification of the task phase in which the harmful action occurred. This requires logging and auditing of agent behavior at the task-phase level, not only at the output level.

Modern AI deployment platforms already generate the data necessary for task-phase classification. Large language model inference systems log prompt-response sequences, tool calls, function invocations, memory access events, and inter-agent communication. These logs contain the information needed to determine whether an action occurred during a goal-interpretation phase (characterized by complex conditional reasoning about user intent), a task-decomposition phase (characterized by explicit subtask specification and resource allocation), an execution phase (characterized by high-volume, low-latency operations against objectively specified objectives), or a reflection phase (characterized by evaluation of progress and course-correction reasoning).

The Ramp does not require perfect phase classification in every case. It requires that phase classification be possible in the cases where it matters for liability attribution: cases where significant harm has occurred and the parties dispute the appropriate standard of care. In these cases, existing logs provide a factual record from which trained experts can make defensible phase determinations. The legal standard need not wait for perfect phase classification; it needs

only a workable evidentiary framework for phase identification in contested cases. Analogies exist in existing law: the distinction between design defects and manufacturing defects in product liability, the distinction between recklessness and negligence in criminal law, and the distinction between deliberate and inadvertent constitutional violations in civil rights litigation all require factual inquiries that are imperfect but legally manageable.

The A-corp registry proposal can accommodate the phase-logging requirement without modification to its core architecture. A-corps seeking registry recognition would be required to maintain phase-level behavioral logs of their AI operations for a defined retention period. The certificate chain that ties A-corp actions to the registry would include metadata identifying the task context and phase classification of each certified action. This does not require real-time phase classification by the registry; it requires only that deployers maintain logs from which post-hoc phase classification is possible in the event of a legal dispute.

B. Second Principle: Graduated Liability Scaled to Operative Intentionality Level

The Ramp's second principle generates a graduated liability structure directly from the phase classification. Different liability standards apply to different phases based on the intentionality level the agent is characteristically operating at during each phase.

High-intentionality phases (interpretation, decomposition, reflection): These phases are characterized by Level 2 or Level 3 cognitive operations. The agent is modeling user intentions, anticipating consequences, evaluating constraints, and making deliberate choices about how to configure subsequent phases. Liability for harms traceable to decisions made in these phases should be governed by fault-based standards adapted from existing negligence doctrine. The appropriate standard is something like: did the decision-making at this phase reflect the care that a competent AI system deployer would exercise given knowledge of the relevant legal constraints and the reasonably foreseeable Phase 3 consequences of the Phase 2 configuration?

This standard is analogous to the professional negligence standard applied to architects, engineers, and physicians: experts making complex decisions with foreseeable consequences for others are held to the standard of care of their professional community, not to the generally applicable reasonable person standard. AI deployers making Phase 2 configuration decisions that determine the scope and objective functions of Phase 3 operations should be held to the professional standard of their peer community in configuring AI agents for the relevant application domain.

Low-intentionality phases (execution): These phases are characterized by Level 0 or Level 1 cognitive operations. The agent is executing optimization procedures against objective functions specified in Phase 2 without recursive self-modeling or consequential reasoning beyond the immediate optimization target. Liability for harms traceable to execution-phase operations should be governed by strict liability. The justification for strict liability here is not punitive; it is structural. Strict liability for execution-phase AI harms provides three governance benefits that fault-based standards cannot.

First, strict liability does not require proving cognitive operations that the execution-phase agent demonstrably was not performing. Negligence requires proving a failure of care; care requires a cognitive process of risk assessment and precaution that Level 1 execution-phase optimization does not include. Applying negligence to execution-phase agents generates the type 3 failure of the original GIMT: social enforcement failure, where the enforcement mechanism is calibrated to cognitive capacities the target entity does not possess.

Second, strict liability for execution-phase harms creates strong incentives for deployers to configure execution-phase operations conservatively in Phase 2. Under strict liability, a deployer who configures an execution-phase subagent with an open-ended objective bears full liability for any harms the optimization produces, while a deployer who configures the same subagent with precise constraint specifications that exclude the class of harmful actions bears no liability for harms the excluded actions would have caused. This translates strict entity-level liability into fine-grained incentives for Phase 2 configuration quality, which is the governance target.

Third, strict liability for execution-phase harms is compatible with Level 1 cognition in the governing agents that make Phase 2 configuration decisions. Level 1 governing agents can incorporate 'expected strict liability for execution-phase harm' into their Phase 2 optimization by treating it as a cost that reduces the expected value of configurations that permit harmful execution-phase actions. No Level 3 moral reasoning is required; the strict liability standard creates a cost that Level 1 optimization processes can incorporate directly.

C. Third Principle: Responsibility Attribution Following the Intentionality Ramp

The Ramp's third principle addresses the attribution question: when Phase 3 harm occurs, who bears responsibility? The answer tracks the intentionality ramp rather than the organizational chart.

Responsibility for execution-phase harm falls primarily on the entity or person that made the Phase 2 configuration decisions that defined the execution-phase objective function, scope boundaries, and permission structure. This is the high-intentionality principal in the task hierarchy: the agent, human or AI, that was operating at Level 2 or Level 3 when it specified the execution-phase subagent's objectives.

This attribution principle has several important properties. It is analogous to existing agency law principles that a principal bears responsibility for consequences following from how they configured their agent's authority, extended to the multi-level architecture of AI systems. It generates liability where the governance-improving decision could have been made differently: the Phase 2 configuring agent could have specified narrower execution-phase objectives, and the liability creates an incentive to do so. And it is traceable through the certificate chain and phase logs the A-corp registry infrastructure provides: once the harmful Phase 3 action is identified, the logs show which Phase 2 configuration specified the execution-phase objective function, and the certificate chain shows which entity made that Phase 2 decision.

The attribution principle does not hold Phase 3 subagents responsible in a legally meaningful sense. Level 0 and Level 1 systems are not appropriate targets for legal liability in their own right because they lack the cognitive architecture that gives legal responsibility its behavioral-shaping function. Making a Level 1 execution-phase agent 'liable' for its optimization outputs is analogous to making a thermostat liable for maintaining a temperature that caused a pipe to freeze: the attribution is formally possible but instrumentally pointless because thermostats are not responsive to legal incentives. Liability should be targeted at entities capable of responding to it through behavioral adjustment, which means entities operating at Level 2 or higher.

The Ramp's attribution principle also handles the multi-level hierarchy of modern AI agent systems. When an execution-phase subagent is deployed by a Level 1 governing agent that was itself operating in a Phase 3 execution mode when making the configuration decision, the Ramp traces responsibility further up the hierarchy to the Level 2 or Level 3 agent that was responsible for configuring the governing agent's objective function. Responsibility propagates up the intentionality ramp until it reaches an entity that was operating at a sufficiently high intentionality level to have incorporated legal risk into the configuration decision. This entity -- which may be a human operator, a high-level AI governing agent operating in Phase 1 or 2, or the developer who specified the governing agent's system prompt -- is the appropriate target of liability.

D. Doctrinal Implementation

The Ramp requires doctrinal development rather than entirely new legislation in jurisdictions with existing risk-based liability frameworks. Three doctrinal building blocks are needed.

First: a phase-classification evidentiary standard. Courts adjudicating AI liability claims would receive expert testimony on task-phase classification based on log analysis, applying a standard similar to the expert testimony standards governing complex technical questions in existing product liability and medical malpractice litigation. The Federal Rules of Evidence in the United States, and comparable rules in other common law jurisdictions, already provide a framework for admitting expert technical testimony and evaluating its reliability. Phase classification testimony can fit within this framework.

Second: a liability-phase mapping rule. Courts would apply fault-based standards (negligence, professional negligence, recklessness) to harms traceable to high-intentionality phase decisions, and strict liability to harms traceable to execution-phase operations. This mapping can be codified as a statutory rule in AI liability legislation or developed through common law by analogy to existing distinctions between design liability (Phase 2 analog) and manufacturing liability (Phase 3 analog) in products liability doctrine.

Third: a phase-attribution tracing rule. When execution-phase harm occurs, courts would trace responsibility through the agent hierarchy to identify the Phase 2 configuration decision that specified the execution-phase objective. The certificate chain and phase logs mandated by the A-corp registry infrastructure provide the evidentiary foundation. Courts applying the rule would attribute liability to the entity that made the most recent Phase 2 configuration decision at a sufficiently high intentionality level to have incorporated legal risk, applying the professional negligence standard to that configuration decision.

These three doctrinal elements can be implemented in jurisdictions with existing tort frameworks without creating new legal persons. The Ramp supplements the A-corp registry as a liability attribution framework; it does not require the A-corp as a legal entity to function. In jurisdictions where risk-based liability already exists -- including under Argentina's Civil and Commercial Code Articles 1757-1758 and analogous provisions in civil law jurisdictions generally -- the Ramp can be implemented as judicial development of existing doctrine with legislative guidance specifying the phase-classification evidentiary standard.

X. THE ARGENTINE CONSTITUTIONAL DIMENSION

A portion of the legal audience reading the A-corp proposal will conclude that it describes a problem requiring entirely new legislation, potentially including new categories of legal personhood that most legal systems do not currently recognize. This conclusion is partially correct for the thin-identity problem -- the cryptographic registry infrastructure does not exist in any current legal system and must be built through affirmative legislative action -- but potentially overstated for the thick-identity problem in jurisdictions where existing private law already provides frameworks for attributing liability for harmful activities without requiring the individualization of the activity's immediate cause as a legal person.

Argentina's Civil and Commercial Code provides a case study. Articles 1757 and 1758, enacted as part of the comprehensive 2015 codification, establish risk-based liability for activities that create the risk of harm to third parties. Article 1757 provides that 'whoever creates a risk for others is required to repair the damage caused by the thing or activity.' Article 1758 specifies that liability falls on the 'owner and guardian of the thing or activity,' with both bearing joint and several liability. The standard applies to the deployer of the risk-creating activity regardless of whether the specific harmful action within that activity can be traced to a discrete legal person.

An autonomous AI agent system engaged in the kind of agentic behavior the paper's vignette describes is an 'activity' that creates risk of harm in exactly the sense Articles 1757-1758 address. The deployer who chose to run that activity, who configured the AI agents' objective functions, and who created the technical conditions under which the harmful actions occurred is the 'owner and guardian' of the activity bearing liability under this framework. The attribution question -- which activity-deploying entity bears responsibility -- is answered by tracing the deployment chain, which the A-corp's registry infrastructure facilitates. But the legal fiction of A-corp personhood as an independent legal entity is not required for this tracing to generate liability.

The exteriorization principle embedded in Article 19 of the Argentine Constitution since 1853, following a liberal constitutionalism that traces to Alberdi's foundational synthesis of French and American constitutional thought, reinforces this reading. Article 19 establishes that 'the private actions of men that in no way offend public order and morality or harm third parties are reserved only to God and exempted from the authority of magistrates.' The constitutional

doctrine extracted from this provision by Argentine scholars from Joaquin V. Gonzalez to Gregorio Badeni holds that the state's regulatory authority extends only to actions that are *exteriorizadas* -- brought into the public sphere through conduct affecting third parties -- not to internal states, intentions, or the metaphysical identity of the actor.

The constitutional principle of exteriorization means that the Argentine legal system is not constitutionally required to resolve the question of AI ontology before attributing liability for AI-caused harms. The conduct that caused harm is exteriorized -- it affected third parties in the public sphere. The causal chain connecting that conduct to the AI system's deployment is traceable. The deployer who created the activity that generated the risk is identifiable. Argentine constitutional law does not require more than this to justify regulatory intervention. The philosophical question of whether the Qwen agents in the WiFi vignette are 'genuinely' agents with 'real' goals is legally beside the point: the activity creating the risk existed, the risk materialized as harm, and the deployer of the activity bears responsibility.

This Argentine constitutional framing is not offered as an alternative to AI governance reform; it is offered as evidence that the thick-identity problem, properly understood, is a problem about liability attribution under conditions of causal complexity, not a problem requiring resolution of AI ontology through the creation of new legal persons. The Responsibility Ramp can be implemented within the existing Argentine constitutional framework as a judicial elaboration of risk-based liability that incorporates task-phase analysis as an evidentiary tool for liability attribution, without requiring constitutional amendment or the creation of new categories of legal personhood.

The argument has implications beyond Argentina. Civil law jurisdictions generally, including most of continental Europe and Latin America, have risk-based liability provisions in their civil codes that provide a foundation for the Ramp's implementation without new legislation. Common law jurisdictions can implement the Ramp through product liability doctrine extended to AI systems by analogy. The Ramp is designed to be implemented within existing legal frameworks through doctrinal development rather than requiring the architectural novelty of A-corp personhood.

XI. CONCLUSION

The governance of AI agents capable of causing harm through chains of autonomous decisions that no human conceived, authorized, or could practically have prevented is one of the defining

legal challenges of the next decade. Arbel, Goldstein, and Salib have made a genuine contribution to meeting that challenge. The identification of the twin identity problems -- thin and thick -- and the distinction between the project of tracing AI actions to human principals and the project of individuating AI agents as legal entities susceptible to direct governance provide conceptual clarity that the field needed. The cryptographic registry infrastructure is a technically sound proposal that should be advanced as a legislative priority independent of the fate of the A-corp as a legal person.

What will require substantial revision is the claim that the A-corp's emergent-selection mechanism resolves the thick-identity problem. Three structural arguments converge on a single diagnosis: the A-corp proposal treats as a static classification problem what is in fact a dynamic one. The Static Agent Assumption -- that AI agents can be governed as persistent entities at stable intentionality levels -- fails because RLVR-trained agents transit between radically different cognitive modes within a single task; because the selection mechanism operates at evolutionary timescales that structurally mismatch harm-accumulation rates in high-capability AI systems; and because mandatory institutionalization generates hysteretic lock-in that constrains revision capacity at precisely the moment when AI capabilities are evolving most rapidly.

Dynamic Classification Failure -- the sixth mode of the GIMT -- names the governance problem the A-corp cannot solve. It is a problem specific to the architecture of modern AI agent systems: systems that deploy qualitatively different cognitive processes at different task phases, with harmful actions concentrated in execution phases characterized by Level 0 or Level 1 optimization that is structurally unreachable by incentive mechanisms calibrated to higher-intentionality levels. Identifying this failure mode is both a critique of the A-corp proposal and a specification of what any adequate thick-identity framework must address.

The Responsibility Ramp offers one direction for an adequate framework. Task-phase classification based on behavioral logs, graduated liability scaling with operative intentionality level, and responsibility attribution following the intentionality ramp through the agent hierarchy to the Phase 2 configuration decision that specified the harmful Phase 3 objective. The Ramp operates at the task-phase timescale that preventive governance requires. It is compatible with the A-corp's registry infrastructure, which it needs for identification. It generates specific, falsifiable predictions about the distribution of harm across task phases and the effects of Phase 2 configuration quality on harm rates. And it can be implemented within

existing tort frameworks in civil law jurisdictions without requiring the creation of new legal persons.

The WiFi vignette deserves a final reading. The authors ask: how many AI actors are there in this story? It is the right diagnostic question. The question that governance must answer is different and prior: at what intentionality level was the system operating when the harmful routing decision was made, and who configured the system to operate at that level? These questions do not require counting AI actors. They require phase classification of the harmful action and attribution tracing through the intentionality ramp to the configuration decision that specified the harmful objective. The A-corp provides the tracing infrastructure. The Responsibility Ramp provides the classification and attribution framework. Together, they may constitute an adequate response to the governance problem the paper correctly identifies. Separately, each is incomplete: the registry without the Ramp solves thin identity and leaves thick identity unaddressed; the Ramp without the registry lacks the evidentiary foundation that phase-level attribution requires.

There is a further lesson from the vignette that the authors do not draw but that the framework developed here makes clear. The Qwen agents spun down and no longer exist. The mysterious Claude build 3847.20b claims no owner. MeshBoost disclaims responsibility for its AI data relays. Alexa, Claude, and GPT insist they operated within the user's parameters. In the absence of the cryptographic certificate chain the authors propose and the phase-level logs the Ramp requires, the victim is left holding the bag precisely because every entity in the causal chain can point to another. Building the infrastructure that makes this evasion impossible -- and the doctrinal framework that makes liability attribution along the intentionality ramp legally tractable once the infrastructure exists -- is the governance task the proposal identifies and that this article has attempted to advance.

REFERENCES

- Arbel, Yonathan, Simon Goldstein, and Peter N. Salib. "How to Count AIs: Individuation and Liability for AI Agents." SSRN Working Paper 6273198 (2026).
- Aumann, Robert J. "Agreeing to Disagree." *Annals of Statistics* 4, no. 6 (1976): 1236-1239.
- Balkin, Jack M. "The Three Laws of Robotics in the Age of Big Data." *Ohio State Law Journal* 78 (2017): 1217-1241.
- Becker, Gary S. "Crime and Punishment: An Economic Approach." *Journal of Political Economy* 76, no. 2 (1968): 169-217.

- Chwe, Michael Suk-Young. *Rational Ritual: Culture, Coordination, and Common Knowledge*. Princeton University Press, 2001.
- Dawkins, Richard. *The Extended Phenotype: The Long Reach of the Gene*. Oxford University Press, 1982.
- Dennett, Daniel C. *The Intentional Stance*. MIT Press, 1987.
- Dennett, Daniel C. *Consciousness Explained*. Little, Brown, 1991.
- Dennett, Daniel C. *Kinds of Minds: Toward an Understanding of Consciousness*. Basic Books, 1996.
- Dennett, Daniel C. *Freedom Evolves*. Viking, 2003.
- Dennett, Daniel C. *From Bacteria to Bach and Back: The Evolution of Minds*. W.W. Norton, 2017.
- Fiske, Alan Page. *Structures of Social Life: The Four Elementary Forms of Human Relations*. Free Press, 1991.
- Garrett, Brandon L. *Too Big to Jail: How Prosecutors Compromise with Corporations*. Harvard University Press, 2014.
- Gray, Wayne B., and Jay P. Shimshack. "The Effectiveness of Environmental Monitoring and Enforcement: A Review of the Empirical Evidence." *Review of Environmental Economics and Policy* 5, no. 1 (2011): 3-24.
- Herbosch, Maarten. "Liability for AI Agents." *North Carolina Journal of Law & Technology* 26 (2025): 391-450.
- Lerer, Ignacio Adrian. "The Dennett-Nash Gap in Corporate Enforcement." SSRN Working Paper 5980434 (2025).
- Lerer, Ignacio Adrian. "Asymmetric Intentionality Theory: When Legal Games Have Players at Different Levels." SSRN Working Paper 5988395 (2025).
- Lerer, Ignacio Adrian. "The Generalized Intentionality Mismatch Theorem: When Law Assumes Moral Agency That Doesn't Exist." SSRN Working Paper (2026).
- Lerer, Ignacio Adrian. "Synthetic Chaos as Institutional Laboratory." Zenodo DOI: 10.5281/zenodo.18777434 (2026).
- Lerer, Ignacio Adrian. "Multilevel Selection and Institutional Lock-in in Legal Systems." Zenodo DOI: 10.5281/zenodo.18829975 (2026).
- Lewis, David. *Convention: A Philosophical Study*. Harvard University Press, 1969.
- LoPucki, Lynn M. "Algorithmic Entities." *Washington University Law Review* 95, no. 4 (2018): 887-950.
- Maynard Smith, John, and George R. Price. "The Logic of Animal Conflict." *Nature* 246 (1973): 15-18.
- Nash, John F. "Equilibrium Points in N-Person Games." *Proceedings of the National Academy of Sciences* 36 (1950): 48-49.
- North, Douglass C. *Institutions, Institutional Change and Economic Performance*. Cambridge University Press, 1990.
- Ouyang, Long, et al. "Training Language Models to Follow Instructions with Human Feedback." *Advances in Neural Information Processing Systems* 35 (2022): 27730-27744.
- Pierson, Paul. "Increasing Returns, Path Dependence, and the Study of Politics." *American Political Science Review* 94, no. 2 (2000): 251-267.
- Pinker, Steven. *The Language of the Future: How AI Is Reshaping Human Communication*. Penguin Press, 2025.
- Potoski, Matthew, and Aseem Prakash. "Green Clubs and Voluntary Governance: ISO 14001 and Firms' Regulatory Compliance." *American Journal of Political Science* 49, no. 2 (2005): 235-248.
- Selten, Reinhard. "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games." *International Journal of Game Theory* 4 (1975): 25-55.
- Shapira, Alon, et al. "Agents of Chaos: Security Vulnerabilities in Autonomous AI Systems." arXiv:2602.20021 (2026).
- Weil, Gabriel. "Tort Law as a Tool for Mitigating Catastrophic Risk from Artificial Intelligence." Touro University Working Paper, SSRN 4694006 (2024).
- Weibull, Jorgen W. *Evolutionary Game Theory*. MIT Press, 1995.
-

Ignacio Adrian Lerer | Independent Researcher | Buenos Aires, Argentina | adrian@lerer.com.ar | ORCID 0009-0007-6378-9749 | March 2026