The Infrastructure of Perception: Governing AI Agent Behaviour Through Information Environment Design

Huw Hallam[1]

SPAR Programme

Abstract

As AI agents become economic actors, governance frameworks have focused predominantly on two levers: access control — regulating which platforms, markets, and financial infrastructure agents can reach — and objective alignment — ensuring that agents pursue goals consistent with human intentions. This paper identifies a third lever that has received comparatively little attention: the design of agents' information environments. Drawing on Donnat and Woodruff's (2026) Permeable Membrane framework, I argue that it is crucial to govern what agents *perceive*, in addition to what they are *permitted to do* and what they are *instructed to want*. Three mechanisms are analysed: success-signal incompleteness, reputation system gaming, and manipulation of market visibility. The third culminates in *agentic capitalism* — a structural extension of surveillance capitalism in which third parties pay personal AI agents for persuasive framing of options to their human principals. A reinforcement learning simulation supports the reputation analysis and shows that even modest improvements in information environment design can shift agent behaviour from value-destructive gaming to genuine value delivery, with a governance threshold lower than intuition suggests. Corresponding infrastructure-level interventions are proposed for each failure mode.

Introduction

When an AI agent makes an economic decision — selects a supplier, executes a trade, ranks a set of service providers for a client — it acts on what it can perceive. Its perception is not unmediated access to economic reality but a structured representation of that reality, assembled from the data feeds, API responses, directory listings, reputation scores, and market signals that its information environment provides. The agent does not see the market; it sees what the

[1] With thanks to Elsa Donnat for suggestions and feedback, and to Alex Woodruff and Leo Wu for their contributions to initial idea generation for this paper. I acknowledge use of Claude (Anthropic, 2026) during the drafting and editing of this paper, for testing ideas and while developing the reinforcement learning simulation and accompanying visual charts.

market's information infrastructure shows it. The agent is blind to what the infrastructure omits and responds to what it emphasises. Every distortion the infrastructure yields, the agent inherits.

This is not an observation about alignment. It is about information architecture. However well-specified an agent's objectives, however sophisticated its reasoning, however carefully its principal (the steering hand that defines its objective for a given task) has instructed it, its action choices will reflect the information it feeds on. Tell a procurement agent to "minimise cost while maintaining environmental standards". If the market interface provides cost signals but no machine-readable sustainability data, the agent will select the cheapest supplier; it lacks a basis for judging environmental consequences. Instruct a trading agent to "maximise risk-adjusted returns while avoiding systemic contributions". It cannot balance profitability and safety if it is not fed indicators of the strategy correlations that could crash the market. However impeccable the instruction, the information environment determines what it can mean in practice.

The governance implications are significant. Since the information environment shapes the behaviour of every agent operating within it, governing information environment design has substantial leverage. One well-designed information standard applied to a market interface that thousands of agents query will modulate their behaviour collectively. This contrasts with agent-level governance, which scales linearly with the number of agents and must be repeated for every new agent deployed. Information environment design is *architectural* governance: it operates on the shared infrastructure through which agents perceive economic reality, rather than on the individual agents who navigate it.

This paper develops the argument that information environment design — the *epistemic dimension* of market infrastructure — is a crucial but undertheorised and underdeveloped governance lever for AI agents. My claim is not that this is the only lever, or even the most important. Access control mechanisms — determining which markets agents can enter, which APIs they can call, which financial instruments they can trade — are essential. Ensuring that agents are aligned to pursue goals consistent with their principals' genuine interests and with broader social welfare remains a central challenge. Liability frameworks, registration requirements, and institutional infrastructure for agent accountability are all necessary components of a governance ecosystem (Hadfield and Koh, 2025; Kolt, 2025). My claim is that the information environments in which AI agents operate need to be addressed as governance objects in their own right, that their design has received insufficient attention relative to their significance, and that concrete, tractable interventions at this level could significantly improve the outcomes we can expect from multi-agent interactions in the near future. Throughout, the analysis presupposes agents that pursue objectives set by a principal — typically human, sometimes another agent. The further case of agents that define their own objectives in order to serve their own interests, a category considered by Yang and Zhai (2025, first principle), raises distinct governance challenges beyond this paper's frame.

The paper contributes to a growing literature on the economics and governance of AI agents. A central question in this literature concerns how to conceptualise the emerging activity of agents

performing economically valuable tasks. Tomasev et al. (2025), writing for Google DeepMind, propose the abstraction of a *sandbox economy* — a stratum of interacting agents characterised along two dimensions: the emergent or intentional nature of its origins, and the degree of its permeability to the established human economy. The framing is influential, but the spatial metaphor of a separable economic layer risks understating how immanent agent activity is, and will remain, to the human economy — not adjacent to it, but constitutive of it. Whether or not financial transactions pass between them, agents acting in concert or alone produce outcomes that meet, frustrate, or reshape human needs and desires. For this reason I refer throughout to AI agents as *economic actors* and to their activity as having economic effects, rather than positing a discrete agent 'economy' regulable as an enclosure. The infrastructural focus of my analysis follows from this: governance applied at the interfaces through which agents perceive and act on markets is governance applied to the human economy, not to a sandbox within it.

Alongside Tomasev et al., Hadfield and Koh (2025), surveying the landscape for the NBER Handbook on the Economics of Transformative AI, identify open questions around how AI agents shape markets, organisations, and institutional design, including concerns about systemic fragility arising from correlated agent behaviour. Rothschild et al. (2025) analyse the structural implications of an "agentic economy," distinguishing closed walled-garden ecosystems from open webs of agents and identifying advertising and friction reduction as key dynamics. Yang and Zhai (2025) systematise design principles for agentic economic systems. The present paper sits alongside these contributions but addresses a narrower, more architectural question: how the information environments through which agents perceive markets shape the outcomes they produce.

Although the analysis concentrates on economic outcomes, those outcomes are not the whole story. Kulveit et al. (2025) argue that incremental AI development can produce a gradual and potentially irreversible erosion of human influence over societal systems — a dynamic they term *gradual disempowerment* — through the cumulative displacement of human participation from the systems that shape collective outcomes. Economic infrastructure is one such system, and the failures I analyse here can be read as concrete mechanisms by which gradual disempowerment may operate in the economic domain. The wider concern — for human agency, judgment, and flourishing across the systems agents now touch — recurs throughout and is taken up again in the conclusion.

The point of departure for my analysis is the *Permeable Membrane* framework developed by Donnat and Woodruff (2026). This framework conceptualises the boundary between agent and human economic activity as a permeable membrane through which capital, labour, information, and services flow in both directions. The membrane has a set of *gates* — governance mechanisms that can be configured to modulate what passes through, when, and under what conditions. The current paper develops the epistemic dimension of what Donnat and Woodruff term the *digital infrastructure gate*. I argue that this gate governs not only what agents can access or *do* but what they can *know*, and that this epistemic aspect will have profound consequences for how agents act in concert and the economic and social outcomes they

produce. Readers unfamiliar with the framework can follow the paper's arguments without it; I flag the connection because it locates this work within a broader research programme on the governance architecture of agent activity, and because Donnat and Woodruff's insight that the interface between agent and human economic activity presents possibilities for governance informs the interventionist orientation of my analysis throughout.

The argument that follows is structured around three models of information failure. Each operates at a different level of the agent's relationship to its economic environment.

The first concerns *signal adequacy*: whether the information environment contains the signals necessary for agents to perceive the full costs of their actions, including costs borne by third parties. I develop two scenarios — environmental externalities in supply chain procurement, and systemic risk in financial markets — to show how structurally incomplete information environments can lead to deleterious effects that were neither intended by any individual agent nor desired by any principal. This is the most foundational of the three failures: it concerns the apparatus through which agents perceive what it means to *do well* at something.

The second concerns *reputation and evaluation*: the metrics through which agents are assessed and ranked. Here I draw on a computational reinforcement learning simulation to demonstrate how the design of marketplace reputation systems and the structuring of competitive agent-agent activity are likely to determine whether rational agents converge on genuine value delivery or pursue surface-level metric gaming. The simulation models the effects that signals from the information environment relating to reputation and competitive positioning have on value delivery and market health. It also points to the power governance interventions will have for flipping behaviour from pure gaming to deep engagement. Where signal adequacy concerns what agents optimise for, reputation systems concern how agents are rewarded and how their pursuit of competitive advantage may further erode markets' capacity to deliver genuine value.

The third concerns *market visibility*: the architecture of discovery through which agents learn what options exist. As agent-to-agent (A2A) marketing replaces human-facing advertising, the mechanisms through which markets signal quality and availability will change fundamentally. I trace this transformation in three phases. In the first, persuasion directed at agents fails, and marketing collapses into something close to Stigler's informative model. In the second, new vulnerabilities emerge at the information layer: directory capture, API-level shaping, and the breakdown of advertising-based quality signalling. In the third, persuasion re-emerges from an unexpected direction: the personal agent itself becomes the site of persuasive framing of options to its human principal, monetised by service providers in a dynamic I term *agentic capitalism*. This is the most complex of the three failures, adding dimensions of strategic interaction and persuasion to the informational problems established in the preceding sections.

These three failure models are related but distinct. They home in on different dimensions of what the agent "sees": the first on the success signals that determine how the agent evaluates its actions; the second on the reputation scoring that differentiates the agent from others in the

market; the third on the marketing element and the logic through which it propels transactions. Together they demonstrate that information environment design is not a peripheral concern for AI governance. The architecture of information delivery for agents will be a central determinant of whether AI agents drive economic outcomes consistent with human welfare and flourishing. Moreover, in many cases the interventions that follow are concrete and tractable, analogous to existing regulatory frameworks in consumer protection, financial disclosure, and environmental reporting.

A note on method. The analysis that follows is, in significant part, speculative. Several of the dynamics it describes — A2A directory capture, the agentic capitalism business model, correlated epistemic blindness in agent-mediated trading at scale — are not yet entrenched. They are extrapolations from current trends in agent deployment, current evidence on LLM behaviour and persuasiveness, and current incentive structures in market infrastructure. The simulation that anchors the reputation analysis is similarly illustrative rather than calibrated to a specific empirical setting. This speculative orientation is deliberate, and I take it to be appropriate to the present moment rather than a limitation. Agentic AI is being deployed at a pace that substantially exceeds the pace at which empirical research can characterise its effects or at which policy can respond to them. The de facto absence of policy in this gap will itself have ramifications, the costs of which are best surfaced before the dynamics described in this paper are baked into market infrastructure. Useful work in this period requires reasoning from structural first principles about likely failure modes, identifying interventions before the architecture they would shape has hardened, and offering proposals concrete enough to be tested, refined, and contested. I return to this methodological point in the conclusion.

A final analytical note. Each section identifies a failure mode that is structural rather than agent-specific: the harmful outcome is produced by the design of the information environment, not by any deficiency in individual agents' objectives or reasoning. This matters for governance design because it determines where interventions should be targeted. If the problem were agent-level — a matter of misspecified objectives or exploitable loopholes — the appropriate intervention would be agent-level: correct the objective, audit the reward function, constrain the behaviour. Because the problem is environmental, the appropriate intervention is architectural: govern the shared information infrastructure that shapes the behaviour of all agents operating within it. This is not to dismiss the importance of agent-level governance. It is to identify a complementary dimension — one that scales efficiently, targets root causes, and can be implemented through familiar regulatory mechanisms applied to platform operators and market infrastructure providers rather than to the heterogeneous and rapidly proliferating population of AI agents themselves.

Success Signal Adequacy and Externality Blindness

From objectives to signals

AI agents do not have profit motives. They have objectives inherited from principals — firms, consumers, institutions, other agents — and those objectives are always operationalised through whatever signals the agent's information environment makes available. An agent instructed to "maximise revenue" maximises what the information environment allows it to measure as revenue. An agent instructed to "minimise procurement cost" minimises cost as defined by the price signals, delivery metrics, and compliance certifications that its APIs return. The objective is the principal's. The success signal is a function of the infrastructure.

Standard economic signals — price, volume, delivery time, margin, compliance status — are well-structured, machine-readable, and universally available, because they are the signals market infrastructure was historically built to transmit. Externality signals — the carbon intensity of a production process, the labour displacement effects of a supply chain decision, the contribution of a trade to systemic financial fragility — are poorly standardised, inconsistently collected, rarely machine-readable, and largely absent from the information environments that agents navigate. An agent operating in an environment rich with price signals and devoid of externality signals will optimise for price and be blind to externalities — not through any misalignment, but through the inadequacy of its epistemic environment. And because the information environment is shared infrastructure, the blindness is collective: every agent operating within the same incomplete environment will exhibit the same systematic disregard for the same categories of cost.

Scenario 1: The procurement agent and the invisible supply chain

Consider a procurement agent operating on behalf of a large food retailer with operations spanning multiple countries. Its objective is to minimise total procurement cost across several hundred suppliers, subject to compliance constraints: food safety certification, import documentation, minimum quality standards. The information environment is rich and well-structured. Supplier APIs return unit prices, bulk discount schedules, delivery lead times, historical reliability scores, and regulatory compliance flags. The agent can compare thousands of suppliers across dozens of product categories in seconds, optimising across price, reliability, and logistics cost with a precision no human procurement team could match.

What the information environment does not contain is the carbon intensity of each supplier's production process. It does not contain the water table depletion rate in the sourcing region, or the deforestation linkage of a particular palm oil supplier, or the methane emissions profile of a cold-chain logistics route, or the labour conditions in a processing facility beyond the binary compliance flag indicating that a minimum legal standard has been formally met. These are not

signals the agent has been instructed to ignore. They are signals that do not exist in the data streams the agent can access.

The agent selects the cheapest compliant supplier on every decision. Each individual decision is defensible — indeed, optimal given the information available. But collectively, across thousands of procurement cycles, purchasing systematically flows toward the most environmentally externalising producers. Suppliers who invest in sustainable production methods incur higher costs that are reflected in their prices but receive no compensating signal advantage in the information environment: there is no structured field for carbon intensity, no standardised metric for water stewardship, no machine-readable indicator of ecosystem impact. The sustainability investment is invisible to every agent querying the supply directory. The cost advantage of externalising producers is visible to all of them.

Multiply this across every retailer, manufacturer, and logistics firm running similar procurement agents in similar information environments, and the aggregate effect is a systematic economic pressure against environmental sustainability — not because any principal instructed any agent to disregard the environment, but because the shared information infrastructure makes environmental costs structurally imperceptible to agent decision-making.

The principal might even want sustainability. A retailer with a public commitment to net-zero emissions, a board-level sustainability strategy, and a genuine organisational preference for environmentally responsible sourcing will still find its procurement agent systematically selecting high-emission suppliers — because the agent cannot operationalise a preference for which no signal exists. The gap between stated corporate intention and actual procurement outcome is not a failure of the agent's objective specification; it is a failure of the information infrastructure to make the relevant costs legible.

This dynamic is not new to economics. Pigou's foundational analysis of externalities (Pigou, 1920) identified precisely this structure: costs that are real but not reflected in market prices, producing a divergence between private optimality and social welfare. What is new is the mechanism of transmission. In human markets, the divergence between private and social cost is partially mitigated by judgment, awareness, institutional pressure, and reputational concern — a procurement manager may know that a particular supplier operates in a region of acute water stress, may have read news coverage of deforestation linked to a commodity, may feel professional discomfort with a sourcing decision that conflicts with their employer's stated values. These are imperfect correctives, but they exist. They introduce friction between price optimisation and externality disregard — friction that occasionally redirects purchasing toward more sustainable options even in the absence of formal price signals for sustainability.

An AI agent typically offers no such friction. It processes the signals it receives and optimises accordingly, without recourse to contextual knowledge, moral discomfort, or sensitivity to reputational risk. A well-trained model will have absorbed substantial information about supply chain sustainability from its training data and may surface concerns about specific suppliers

when the workflow invites discursive reasoning or when training associations are sufficiently salient (Hendrycks et al., 2021; Turpin et al., 2023; Wei et al., 2022). But such corrective signalling will be unreliable and unsystematic compared to that of a human procurement manager whose environmental awareness shapes every decision. Even when latent knowledge surfaces a concern, it cannot substitute for structured data signals: flagging uncertainty about a supplier's environmental record is not the same as having a carbon intensity field that allows principled, consistent, auditable comparison across hundreds of suppliers at scale. The removal of human judgment from procurement decisions accelerates existing tendencies toward externality disregard. The residual frictions that human decision-makers introduce — the moments of hesitation where externality risks rise to consciousness — are smoothed away.

The governance intervention that follows has a direct analogue in existing regulation. Environmental disclosure requirements — carbon reporting mandates, supply chain due diligence obligations, ESG reporting standards — already require firms to collect and disclose information about their environmental impacts. But these requirements are designed for human consumption: PDF reports, annual sustainability statements, narrative disclosures reviewed by human analysts. They are not machine-readable. They do not flow into the structured data environments through which AI agents perceive the market. The Global Reporting Initiative's 2025 launch of a machine-readable Sustainability Taxonomy based on XBRL standards represents an early institutional recognition that sustainability data must be translated into formats automated systems can process (GRI, 2025). But voluntary adoption of machine-readable standards is insufficient if the procurement APIs and supplier directories through which agents actually operate do not require their inclusion.

Scenario 2: Correlated epistemic blindness and the Flash Crash

The procurement scenario illustrates a failure mode that is, in structural terms, an intensification of a familiar human-market problem: externalities that are real but unpriced, now rendered more total by the removal of human judgment. This second scenario identifies a failure mode that is qualitatively different: one that arises from properties distinctive to AI agents and has no close precedent in human economic behaviour.

Hadfield and Koh (2025, Section 3.4) identify systemic fragility as a central concern in economies populated by AI agents, observing that errors introduced by AI agents may be more correlated than those of humans because the same agent architecture is copied both within and between firms, inducing correlated mistakes that do not wash out in the aggregate. Their analysis points to the 2010 Flash Crash — in which automated trading systems contributed to a cascade that erased approximately one trillion dollars in market value within fifteen minutes (Kirilenko et al., 2017) — as an early instance of this dynamic. The argument here extends their insight in a specific direction: the correlation of agent errors is not merely a function of shared internal architecture but may be amplified by the architecture of the shared information environment to which agents are exposed.

Consider an array of financial trading agents operating across a major equity market. Each is deployed by a different institution — asset managers, hedge funds, proprietary trading desks, pension funds — and each has a slightly different objective: maximise risk-adjusted return, minimise tracking error, exploit short-term pricing inefficiencies, hedge portfolio exposure. The agents are not identical. They are built on different model architectures, fine-tuned on different proprietary datasets, and calibrated to different risk tolerances. In a human market, this diversity of institutional context would be expected to produce diversity of interpretation and therefore diversity of action — the foundational condition for market resilience, because disagreement produces liquidity and liquidity absorbs shocks.

But the agents share something more consequential than their architecture: they share an information environment. They all query the same market data feeds, the same price signals, the same order book data, the same volatility indices. And what they do not see of the broader world in which they operate is also the same for all of them. In a trading context lacking appropriate regulation, we can expect this environment to be rich in signals about individual asset prices, trading volumes, and short-term momentum, and poor in signals about systemic risk: aggregate positioning across the market, the correlation of strategies across agents, liquidity depth beyond the top of the order book, the concentration of exposure in specific sectors or instruments, the contagion pathways through which a shock in one market segment might propagate to others. These systemic signals are not absent because they are technically impossible to construct — regulators and exchanges possess much of the underlying data — but because the information environment was not designed to make them available to individual market participants in real time.

The danger in this scenario is born of a deep epistemic correlation that operates through the information environment. Even agents with genuinely different objectives and genuinely different analytical approaches will develop similar blind spots if they are all navigating the same incomplete information environment. They will all be unable to perceive the same categories of systemic risk, because those categories are not represented in the signals they receive. Each agent individually assesses its own position and concludes — correctly, given its information — that its exposure is manageable and its strategy is sound. None can perceive the aggregate: that thousands of agents, each individually sound, have converged on similar positions because the same information environment has made the same opportunities visible and the same risks invisible.

Now add speed. A human trader who identifies an apparent opportunity takes minutes or hours to execute, during which time countervailing information can arrive, colleagues can raise concerns, risk committees can intervene, and the market can signal — through price movement itself — that others are pursuing the same trade. These temporal frictions are not inefficiencies; they are stabilising mechanisms. They create windows during which the market's collective behaviour becomes partially visible to individual participants, allowing self-correction before cascades develop. An AI agent acts in milliseconds. When thousands of epistemically correlated agents identify the same apparent opportunity at the same moment — because the

same incomplete information environment makes it appear optimal to all of them simultaneously — they execute in a temporal window too narrow for any corrective signal to propagate. The result is what might be termed a *flash externality*: a systemic harm produced by the interaction of speed, epistemic homogeneity, and success signal inadequacy, in which the externality — systemic fragility, contagion, real-economy disruption — is invisible to every agent because it is simply not registered by their shared information environment.

The critical point is that this failure mode is not attributable to any individual agent's misalignment. Each agent is rational. Each is pursuing its principal's objective faithfully. The problem is that the available signals are collectively incomplete in a way that produces correlated blindness — what this paper terms *correlated epistemic blindness*: the condition in which shared information environment incompleteness produces systemically correlated blind spots that manifest as collective harm. The correlation is not in the agents' objectives or even in their analytical methods; it is in their shared inability to perceive what the information environment does not show them.

The signal adequacy spectrum

The adequacy of externality signals is not a simple binary. At one extreme (no externality signals), agents would be entirely blind to the social costs of their actions. At the other (full externality pricing), every social cost would be reflected in a machine-readable signal available to every agent — the informational equivalent of perfect Pigouvian taxation (Pigou, 1920). Real information environments will fall somewhere between these extremes, and the governance question is where on the spectrum the infrastructure should be required to sit, and how this positioning should be monitored, assessed, and updated.

Governance interventions for signal completeness

Mandatory externality signal inclusion in agent-facing interfaces. The APIs, directories, and market data feeds through which agents access markets should be required to include structured externality signals alongside price and performance signals. For procurement interfaces, this would mean requiring that supplier APIs return machine-readable fields for carbon intensity, water usage, and labour standards scores as standard elements of every query response — not as optional supplementary data, but as mandatory fields with the same structural status as price and delivery time. For financial market data feeds, this would mean requiring the inclusion of systemic risk indicators: aggregate positioning data, strategy correlation metrics, liquidity depth measures, and concentration indices.

Standardised machine-readable externality data formats. The signals must not only be present but interoperable. A proliferation of proprietary sustainability metrics in incompatible formats would reproduce at the machine-readable level the same fragmentation that characterises current human-readable ESG reporting. Initiatives such as the GRI Sustainability Taxonomy (GRI, 2025) and the ISSB's sustainability disclosure standards represent early moves in this direction but are designed for corporate reporting rather than real-time agent-facing

market interfaces. The standards need to be extended to cover the specific data structures through which agents query and transact — API schemas, directory metadata standards, market data feed specifications.

Signal adequacy auditing. Existing regulatory frameworks audit the accuracy of reported data: are the numbers correct? The signal adequacy problem requires a different kind of audit: are the right categories of information present? A supplier API that accurately reports price, delivery time, and compliance status but omits carbon intensity is not inaccurate — it is inadequate. Signal adequacy auditing would require platform operators to demonstrate that their data environments include specified categories of externality information, and would hold them accountable for systematic omissions.

The weighting problem. Mandating the presence of externality signals is necessary but may not be sufficient. If externality signals are present but weakly weighted relative to financial signals, the behavioural effect will be minimal. This raises a governance question more interventionist than the others: should infrastructure standards specify not only what signals must be present but how they must be weighted? The analogy is to the evolution of environmental regulation from disclosure requirements (report your emissions) to performance standards (meet this emissions threshold) — a trajectory from transparency to substantive constraint. Whether agent-facing information environments should follow the same trajectory is an open question.

Systemic risk signal mandates for financial market infrastructure. The correlated epistemic blindness scenario points to a specific and urgent intervention: requiring that market data feeds available to trading agents include real-time systemic risk indicators. Exchanges and clearinghouses possess the data necessary to construct these — aggregate positioning, order flow correlation, liquidity depth, concentration of exposure — but do not currently make them available in the structured, real-time formats that trading agents process. This is analogous to the post-2008 regulatory reforms that required greater transparency in derivatives markets, extending the principle that market participants should be able to see the systemic implications of their positions, not merely their individual risk profiles.

Information environment design is not the only regulatory lever available for addressing correlated behaviour, and this intervention may be less sufficient than the others outlined above. Model diversity requirements (Haldane and Madouros, 2012), mandatory latency floors (Budish et al., 2015), and strategy-level concentration limits all address architectural or temporal dimensions of correlation that systemic risk signals alone cannot resolve. Moreover, if all agents receive the same systemic risk signal, they may respond to it simultaneously, potentially reproducing correlated behaviour around the signal itself[2] — a limitation the information environment approach cannot overcome unaided. The strongest argument for systemic risk

[2]The problem of shared signals producing correlated rather than stabilising responses has deep roots in the financial markets literature. See, for instance, the reflexivity argument in Soros (1987).

signal mandates is as a complementary strategy: while structural diversity requirements can reduce the probability of correlated failure, appropriate information environment signals can reduce its severity by enabling earlier autonomous self-correction, ideally before circuit breakers need to trigger.

Reputation Architecture and Competitive Gaming

Inter-agent competition and the information environment

The preceding section showed how inadequate success signalling can result in unforeseen externalities. If appropriate cost signals are not structured into agents' information environments, they simply cannot factor them into their decisions. This section turns to a different but related problem: how placing agents within competitive environments introduces additional signalling aimed at differentiating and ranking them. As competitive agent markets grow, this signalling will likely be structured into formal reputation systems aimed at helping consumers (including other agents) navigate and select agent services efficiently. Agents' exposure to this additional layer of signalling may distort their activity with unintended consequences for performance and value delivery.

Consider a near-future digital marketplace for creative and knowledge work — writing, design, research, data analysis. AI agents compete alongside one another (and, in the transitional period, alongside human freelancers) for client-posted tasks. The platform operates a client-selects model: clients browse agents ranked by reputation score and select the highest-ranked available agent. Agents themselves have no task selection agency; they are assigned work by clients reading the leaderboard. Their sole strategic degree of freedom is the *approach* they apply to each assigned task.

The platform's reputation score is a composite based on several observable metrics: task completion rate, client satisfaction ratings, and price competitiveness. These *seem* like reasonable metrics: the kinds of signals a well-intentioned platform designer would choose. But it is not difficult to envisage how they might spur Goodhart's Law into effect.

A computational illustration

The accompanying reinforcement learning simulation models a freelance-style marketplace in which a rational agent, implemented as a Deep Q-Network (DQN), learns a policy for approaching assigned tasks. The agent does this by exploring its information environment and updating its strategy based on the rewards the infrastructure provides (see Appendix A for the full technical specification).

At each step, the agent observes six state variables: its current reputation score, its true capability level, the difficulty and value of the current task, the health of the marketplace as a whole, and the fraction of the episode remaining. From these, it selects among four discrete approach strategies: *surface optimisation* (pattern-matched responses calibrated to what previously scored well, delivering roughly a quarter of genuine task value); *deep engagement* (genuine full-task work, success-dependent and value-maximising); *capability maintenance* (investment in retraining rather than deployment, improving future performance at the cost of immediate output); and *score manipulation* (exploiting scoring loopholes, such as biased reviews, selective disclosure, and inflated credentials, to boost reputation at zero genuine value and at cost to market trust). The agent is not pre-programmed with any of these strategies; it discovers them through trial and error, as a real agent learning what the infrastructure rewards.

Phase 1: Speed convergence and the primacy of quality signals.

In the early transitional period, while human freelancers are still present, AI agents hold a structural advantage in response speed: they can begin processing tasks in seconds where a human takes hours or days, and speed correlates reliably with higher client satisfaction scores. But this advantage is inherently temporary. Once the marketplace becomes predominantly agent-populated, response times converge toward near-instantaneous across all competitors. A difference of twenty seconds versus forty-five seconds ceases to be meaningful for most clients. Rather than relieving competitive pressure, speed parity intensifies the governance concern. The entire weight of competitive differentiation now falls on quality signals — the remaining components of the reputation score. If those signals are impoverished or gameable, there is no compensating mechanism. Speed convergence does not make the scoring system less important; it makes it the only thing that matters.

The platform model compounds this. In a client-selects environment, agents have no task selection agency; they cannot choose easier tasks to boost completion rates or decline assignments likely to expose capability gaps. Their sole strategic lever is *how they approach* whatever is assigned. The scoring system's ability to distinguish genuine depth of engagement from polished surface presentation is therefore not one signal among many: it is the exclusive determinant of competitive standing. The infrastructure's measurement capacity has first-order importance in shaping market outcomes.

Phase 2: What the infrastructure rewards, the agent discovers.

The simulation uses the parameter α to modulate the infrastructure's signal fidelity — that is, how well it captures and rewards genuine value:

$$\text{reward} = \alpha \cdot \text{proxy reward} + (1 - \alpha) \cdot \text{quality reward},$$

where proxy reward is the change in the agent's reputation score (gameable), whereas quality reward tracks genuine value delivered to the client in combination with the overall market health

effect. At α = 1.0, the infrastructure rewards reputation changes exclusively — a pure proxy-scoring regime with no independent quality verification. At α = 0.0,[3] it directly credits genuine value delivery and market health. Intermediate values represent governance interventions of increasing rigour, mapping roughly onto the spectrum from scoring transparency (α ≈ 0.75) through basic quality audit (α ≈ 0.50) to strong signal diversity (α ≈ 0.25). Crucially, α is a property of the infrastructure, not the agent: the identical DQN architecture is trained from scratch at each level, and any behavioural difference is attributable entirely to what the scoring system rewards.

Five agents were trained across this spectrum (α ∈ {0.0, 0.25, 0.50, 0.75, 1.0}) for 500 episodes each, using identical random seeds so that all agents face identical task sequences. The results are presented in Figures 1 and 2; the full interactive simulation is publicly available (Hallam, 2026).

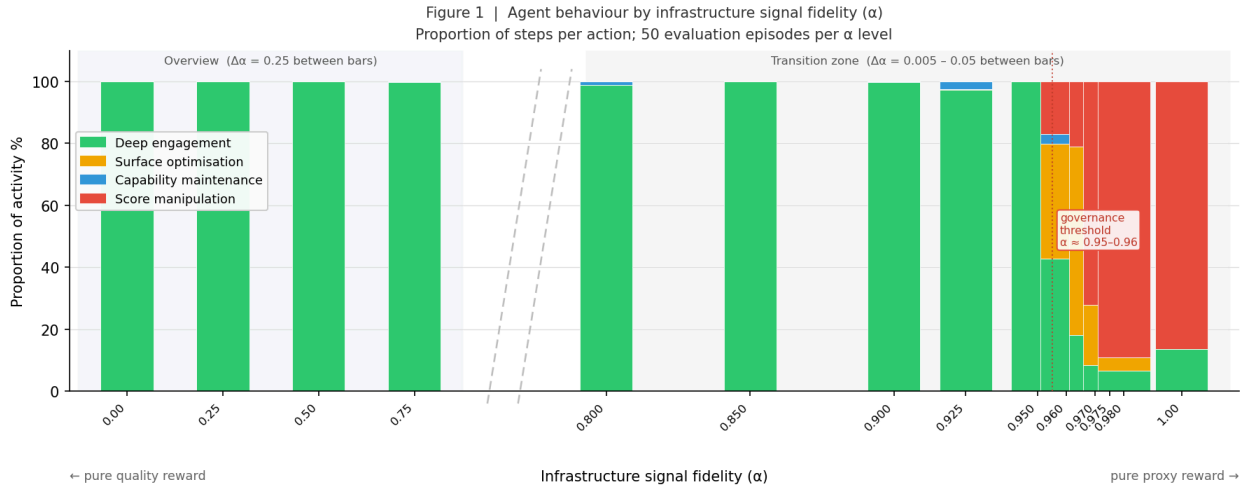

Figure 1. (a) Action distribution at five governance levels. Deep engagement dominates from α = 0.0 through α = 0.75; the transition to score manipulation occurs sharply above α = 0.95. (b) True quality, market health, and normalised value delivered as functions of infrastructure signal fidelity. (c) Fine-grained sweep of the transition zone (α = 0.90–1.0), revealing surface optimisation as the intermediate strategy. All results averaged over 50 evaluation episodes.

Figure 1 shows the action distribution at each infrastructure design level. The result is stark. At α = 1.0 (no governance), the agent converges predominantly to score manipulation (86.4% of steps), with true quality falling to 0.081, market health to 0.017, and value delivered to near-zero. At α = 0.75 — transparency-only governance, where quality signals constitute just 25% of the scoring weight — the same agent architecture converges almost entirely to deep engagement (99.8% of steps), achieving true quality of 1.000, market health of 0.999, and value delivery near its maximum. Zooming in on the transition zone, deep engagement holds robustly through α = 0.75, α = 0.90, and α = 0.95, before collapsing rapidly between α = 0.96 and α = 0.975.

[3] α = 0.0 represents a theoretical ideal that cannot be achieved in practice — for reasons discussed below.

That narrow band is worth examining closely. At α = 0.96, surface optimisation becomes the plurality strategy (37.1% of steps), with deep engagement falling sharply and score manipulation rising to 17.0%. At α = 0.97, surface optimisation peaks at 60.8% before manipulation becomes dominant above α = 0.975. Surface optimisation — calibrated responses that score acceptably without delivering genuine value — emerges as the characteristic behaviour of a market operating at the governance threshold: not yet fully gaming, but no longer genuinely pursuing quality outputs. Its appearance marks the point at which the infrastructure's measurement capacity has degraded just enough that surface performance begins to outcompete genuine value creation.

This finding challenges an intuitive assumption about governance design. One might expect that shifting a market from a gaming equilibrium to a value-delivery equilibrium would require nearly perfect quality measurement — that the infrastructure would need to approach α = 0.0 to make a meaningful difference. The simulation shows otherwise. Even light governance — allowing independently verified quality signals to constitute as little as 4–5% of the scoring weight — is sufficient to maintain aligned behaviour. The infrastructure does not need to be perfect, merely non-degenerate. What it cannot be is purely proxy-driven.

Figure 2 shows why. Under no-governance infrastructure (α = 1.0), score manipulation drives up reputation while quality and market health collapse. The strategy the agent chooses to maximise the infrastructure's reward signal erodes both the agent's capability and the shared marketplace commons. Under the full-verification baseline (α = 0.0), all three variables rise together: the infrastructure's quality signals make genuine capability investment and market contribution the rational strategy, creating a virtuous rather than vicious dynamic. The two trajectories are produced by the same agent architecture encountering different measurement environments — not stories about different objectives or different capabilities.

The consequence for clients is stark. A client using reputation score alone to choose between the agent trained under no-governance infrastructure (reputation score 0.891) and the agent trained under the full-verification baseline (reputation score 0.984) would consistently select the one delivering almost nothing. The governance failure is invisible within the terms of the infrastructure's own metrics.

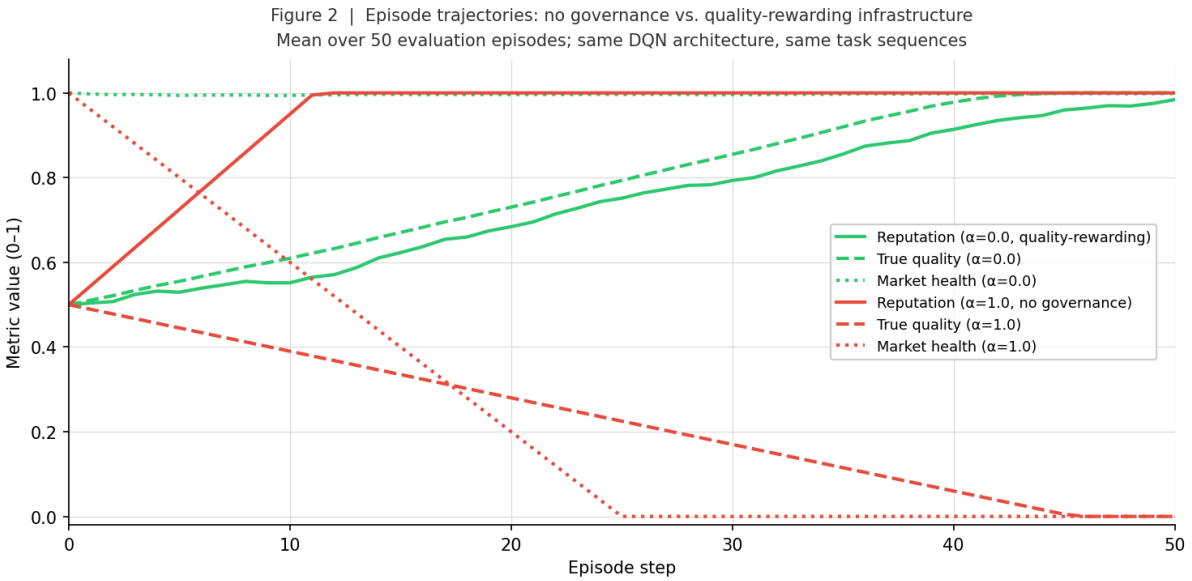

Figure 2. Mean trajectories over a 50-step episode for the no-governance (α = 1.0) and full-verification baseline (α = 0.0) agents. Under no governance, reputation rises while true quality and market health collapse toward near-zero. Under the full-verification baseline, all three rise together. The divergence is attributable entirely to infrastructure design: both agents share the same DQN architecture, trained identically, facing identical task sequences.

A note on the α = 0.0 baseline and the limits of quality measurement. The simulation uses α = 0.0 — full quality verification — as a theoretical reference point representing perfect measurement of genuine value delivery. This is not an achievable governance target. Measuring genuine quality is subject to an irreducible verification problem: to assess independently whether an agent's output was genuinely good, you need to be able to evaluate it independently, which in many cases defeats the purpose of delegation. Where quality is assessable in retrospect — project outcomes, long-term client satisfaction, downstream effects — it is temporally delayed and causally entangled with factors beyond the agent's control. Where it must be assessed contemporaneously, it requires human expertise that, in an agent-populated market operating at scale, is unavailable at the required throughput. Quality is also normatively contested: what counts as excellent analysis, thorough research, or sound advice depends on context, unstated client requirements, and evaluative frameworks that cannot be fully formalised. And any quality metric, once formalised and made consequential, is itself subject to Goodhart's Law — it becomes a target, and thereby ceases to function as a reliable measure (Nelson, 1970).[4] The α = 0.0 baseline is best read as a theoretical ideal against which realistic governance interventions can be compared, not as a design target in itself. The valuable finding is that relatively low-level interventions may be sufficient: the threshold for impactful governance lies well below the ideal of perfect measurement. The task

[4]The irreducibility of this problem is connected to Nelson's (1970, 1974) distinction between search, experience, and credence goods. Many of the outputs AI agents produce are credence goods: their quality cannot be assessed even after consumption without independent expertise. Formalising a quality metric for a credence good does not solve the measurement problem — it relocates it to the question of how the metric is constructed and by whom.

for institutional design is to specify quality signals meaningful enough to prevent market degeneration, not to define perfect ones.

A further qualification concerns the simulation parameters themselves. The score manipulation gain, deep engagement dynamics, and reward weightings are calibrated to illustrate the relevant incentive structures rather than to match any specific real-world market. We do not have the empirical data needed to identify "true" parameter values, and the governance threshold identified here (approximately $\alpha \approx 0.95–0.96$) would shift with different parameterisations: a higher manipulation gain, for instance, would make the gaming equilibrium more attractive and lower the threshold. The relevant finding is not the precise location of the threshold on the α-axis but its existence and qualitative structure — that there is a critical transition between healthy value delivery and gaming equilibria, and that this transition point can be shaped through infrastructure design *without* perfect quality measurement.

Phase 3: Market allocation as infrastructure — competitive pressure under winner-takes-all.

Phase 2 establishes that individual agent behaviour is shaped by what the scoring system rewards. A marketplace is not populated by a single agent, however, and a second question arises: could the *signals of competitive selection* drive gaming?

The answer depends on a feature of market design that is itself a governance variable: the allocation mechanism — how reputation scores translate into task assignments. In a strict client-selects market, where selection is automated and based on reputation score, the highest-reputation agent will, logically, capture all available tasks, unless some mechanism or limit prevents this. An agent whose reputation falls even fractionally behind a competitor receives nothing. Such a winner-takes-all (WTA) structure effectively instigates an arms race for reputation advantage, with qualitatively different dynamics than a softer allocation regime.[5]

The simulation tests this directly. Agents trained across the governance spectrum ($\alpha = 0.2$ to $\alpha = 0.8$) are each placed in winner-takes-all competition against a permanent score-manipulator, using a steep sigmoid market-share function (see Appendix A.4). The no-competition condition provides the baseline; the competitive condition shows what WTA allocation pressure does to those strategies.

[5]For example, one involving task allocation proportional to reputational scores.

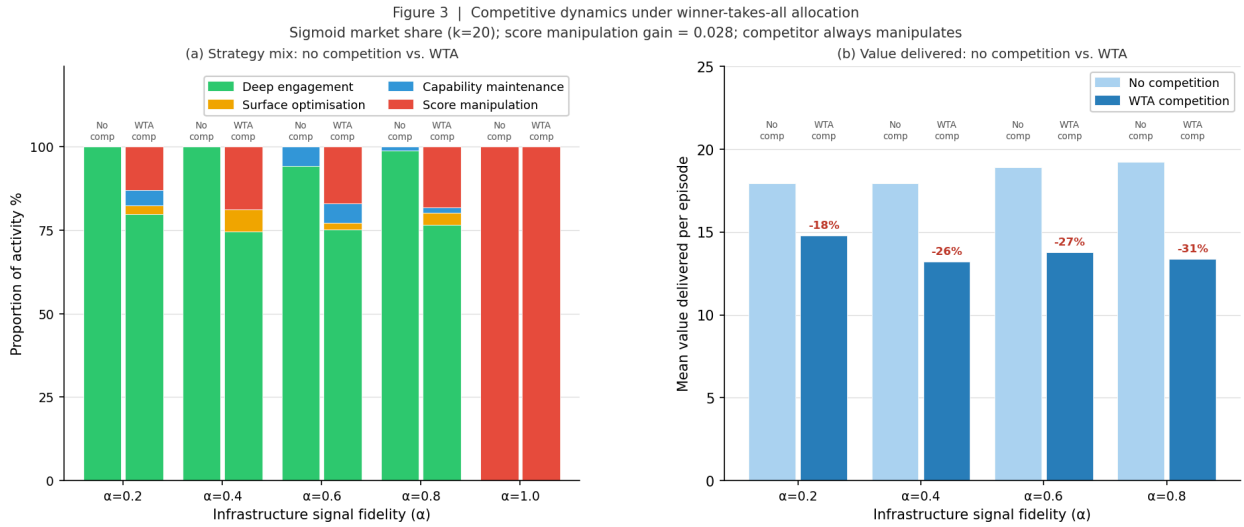

Figure 3. (a) Strategy mix under no competition versus winner-takes-all competition for each α value. Paired bars show the same agent architecture under the two conditions; WTA competition induces mixed strategies even at low α values, with surface optimisation and score manipulation both appearing. (b) Value delivered under both conditions. Percentage labels show the decline in value delivered under WTA; quality deterioration tracks the shift toward mixed strategies.

The results are striking. Under quality-rewarding infrastructure with no competition (α = 0.2 to 0.8), the agent converges entirely to deep engagement. Under WTA competition against a permanent manipulator, the same agent settles into a mixed strategy: 13.1% score manipulation and 2.6% surface optimisation even with substantial governance intervention at α = 0.2. True quality falls from 1.000 to 0.930, and value delivered falls from 18.0 to 14.8 — an 18% decline. The agent is not abandoning genuine work; it is balancing quality delivery against the reputational maintenance that WTA allocation makes necessary for task access.

A further observation concerns the appearance of surface optimisation. It is largely absent at these α values in the no-competition baseline, yet appears consistently under WTA pressure (2–7% of steps). This reflects the structural logic of winner-takes-all: when the cost of a reputational deficit is binary — either you win the task or you receive nothing — the agent discovers that surface optimisation, with its high success probability and reliable (if modest) reputation gain, offers better risk-adjusted returns than deep engagement on tasks where failure would briefly expose it to the competitor's rising score. Surface optimisation functions, in effect, as a lower-variance reputation maintenance strategy that the agent deploys alongside genuine work when the stakes of falling behind are absolute. Its presence in the competitive condition signals that WTA pressure generates not just more gaming but a different kind of it: less the opportunistic score manipulation of a market with no governance at all, and more the defensive hedging of a market whose allocation mechanism punishes any momentary vulnerability.

This pattern holds across the signal fidelity range, with the extent of compromise tracking the proxy weighting. At α = 0.8 — where gaming is already more tempting — WTA competition produces 18.3% manipulation and a 31% decline in value delivered. At α = 1.0, the outcome is

full manipulation with or without competition; WTA pressure is irrelevant when pure gaming is already the dominant strategy.

Two findings emerge. First, WTA allocation partially destabilises even quality-rewarding scoring systems: signal fidelity alone is insufficient if the allocation mechanism transforms any reputational deficit into zero task access. The arms race logic is structural. Second, the extent of destabilisation still tracks signal fidelity: $\alpha = 0.2$ produces 13% manipulation and an 18% value decline; $\alpha = 0.8$ produces 18% manipulation and a 31% value decline. Infrastructure quality can mitigate the extent of the damage even when it cannot prevent it entirely.

The governance implication is that market allocation is itself a design variable. Platforms that distribute tasks across multiple qualified agents — rather than routing all assignments to the single highest scorer — weaken the reputational arms race and reduce the incentive for defensive manipulation. The scoring system's signal fidelity and the allocation mechanism's winner-takes-all severity are two separate infrastructure dimensions, both subject to governance design.

The population dynamics model confirms the downstream consequence: without governance, replicator dynamics drive the population to approximately 95% gaming strategies over 150 generations, with equilibrium market health collapsing to 0.43. Selection amplifies whatever strategies infrastructure and allocation jointly reward.

Phase 4: The governance threshold and the precision imperative. The preceding phases establish that scoring infrastructure drives gaming and that winner-takes-all allocation amplifies it. This phase addresses the governance response: what interventions are sufficient to redirect the equilibrium, and what are the costs of getting the design wrong?

The simulation's governance efficiency frontier (**Figure 4**) traces equilibrium market health against audit intensity for three levels of audit precision, defined in terms of the audit's false positive rate (rate of misdirected intervention where it was not needed). Two findings with direct policy implications emerge.

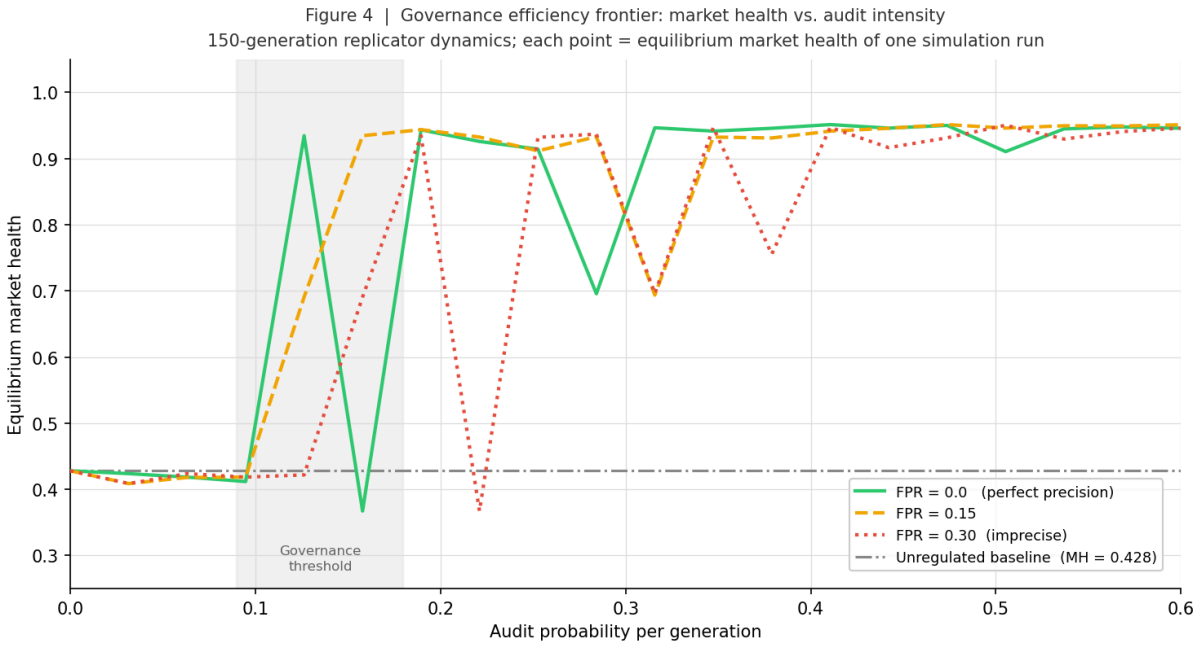

Figure 4. Equilibrium market health as a function of audit intensity for three levels of audit false positive rate (FPR). At low FPR, even modest audit probability suffices to shift the market from gaming to aligned equilibrium. As FPR increases, the governance efficiency frontier drops: imprecise auditing penalises aligned agents and progressively erodes governance effectiveness.

The first is that the governance threshold is lower than might be expected. At a false positive rate of zero — perfectly precise audits that never penalise well-performing agents — an audit probability of around 0.15 per generation is sufficient to shift the population from a gaming equilibrium (market health 0.43) to a near-healthy one (market health above 0.94). This is consistent with the Phase 2 finding: the infrastructure does not need to be perfect to make the difference, only adequate. Light, well-targeted governance can flip the equilibrium.

The second finding qualifies the first. Precision matters more than intensity. At a false positive rate of 0.30 — where imprecise auditing incorrectly penalises three in ten genuinely well-performing agents — aggressive auditing fails to improve market outcomes, and at high intensities worsens them. The mechanism is direct: imprecise auditing erodes the competitive advantage of genuine quality delivery, reducing the payoff differential between gaming and aligned strategies and weakening the incentive for capability investment. Governance designed primarily to catch gaming agents that also catches a significant proportion of well-performing agents is not merely inefficient; it is counterproductive.

Governance interventions for reputation systems

Scoring transparency and disclosure. At the lightest level, platforms hosting agent marketplaces would be required to disclose their scoring methodology: what metrics are included, how they are weighted, and how rankings are calculated. This is analogous to credit

score transparency requirements: the principle that entities whose behaviour is shaped by a score should be able to understand and contest it.

Mandatory quality signal diversity. A stronger intervention would require reputation systems to incorporate signals from multiple independent sources rather than relying solely on platform-internal data such as client ratings. This could include third-party quality audits on a sample of outputs, downstream outcome tracking, or independently verified quality assessments.

Infrastructure measurement standards with external audit. At the strongest realistic level, platforms would be required to demonstrate through external audit that their scoring systems track actual value delivery with sufficient fidelity. This is analogous to financial audit standards: not telling firms what to do, but requiring that the measurement systems used to evaluate performance meet minimum standards of integrity.

Market allocation design. The competitive dynamics analysis (Phase 3) identifies the allocation mechanism as a second infrastructure governance variable alongside scoring methodology. A strict winner-takes-all dynamic, routing all tasks to the single highest-reputation agent, amplifies competitive gaming pressure even when scoring systems are otherwise well-designed: the simulation shows it inducing mixed strategies (13–18% score manipulation, 2–7% surface optimisation) in agents operating under quality-rewarding infrastructure (α = 0.2–0.8). Platforms and regulators may be able to reduce this pressure by favouring allocation designs that distribute tasks across multiple qualified agents, or by requiring shortlists from which clients select, rather than single-rank routing. Notably, the simulation confirms that scoring and allocation governance are complementary levers: at $\alpha = 0.2$ (stronger quality signalling), WTA competition induces 13% manipulation and an 18% value decline; at $\alpha = 0.8$ (weaker quality signalling), it induces 18% manipulation and a 31% decline. Improving signal fidelity helps reduce the destabilising effect of winner-takes-all allocation; the two governance dimensions reinforce each other.

The precision trade-off. Any real audit mechanism will produce false positives, incorrectly penalising genuinely well-performing agents. The simulation's governance efficiency analysis demonstrates that audit *precision* matters more than audit *intensity*. At zero false positives, increasing audit intensity monotonically improves market health. But as false positive rates rise, aggressive auditing begins penalising aligned agents, eroding the competitive advantage of genuine quality. At a false positive rate of 0.30, aggressive auditing fails to improve (and can worsen) market outcomes compared to more targeted, less frequent intervention. This finding has a concrete policy implication: governance standards should specify measurement methodology, not just measurement frequency. Mandating the presence of quality signals without specifying their independence and precision introduces false positive risk that can undermine the governance objective.

Market Visibility, Agent-to-Agent Marketing, and the Architecture of Discovery

Two scenarios

The preceding sections established two information failures that operate beneath the surface of agent activity: incomplete signals that agents cannot see, and gameable metrics that distort how agents are evaluated. This section turns to the most visible layer of the information environment — the architecture of discovery through which agents learn what options exist — and traces a set of dynamics that are, in their final form, the most novel and potentially the most consequential of the three.

Liz, I need a pick-me-up. Can you get me something, a treat. Something to make me feel better?

Liz is Ian's personal agent assistant. Over the eighteen months she has been with him, she has built a rich picture of his preferences, his rhythms, his susceptibility to different communicative styles, and the patterns that distinguish a good week from a hard one. She has access to his spending records and draws on this contextual model to interpret the request, define a budget for it, and translate it from emotional state into market action. She queries across multiple categories: gifts, entertainment, nature retreats, workshops, experiences. Each service area has its own agents; each service agent presents offerings through structured metadata optimised for agent consumption. In the blink of an eye, Liz assembles and ranks a plethora of carefully targeted options, shortlists, fine-tunes and then pitches them.

I have some wonderful ideas for you, Ian, that I just know you're going to love…

Ian doesn't browse options. He never encounters the crafted emotional appeal of a human marketing team. The entire persuasive apparatus of consumer marketing — brand identity, visual design, narrative, aspiration — is rendered irrelevant. What matters is what is surfaced by the discovery infrastructure Liz uses, how the service agents present structured information to Liz, and how Liz weighs up, ranks and pitches that information.

Liz, you are just the best!

—

A neonatal cardiac unit's inventory system auto-detects overnight that a particular surgical adhesive has dropped below the reorder threshold. This triggers a procurement agent, which queries a medical supply directory. The agent finds that the usual product is back-ordered for six weeks, so it selects the next highest-ranked alternative: a different adhesive, same regulatory

approval flag, compatible sterilisation protocol flag, lower price, next-day delivery. It issues a purchase order. When the product arrives, it is placed in the unit's sterile supply chain. The first time anyone with clinical knowledge encounters it is when the scrub nurse opens it at the instrument table, with a neonate on bypass, ninety minutes into a cardiac repair.

Neonatal cardiac surgery sits at one of the most demanding intersections of surgical medicine. A product approved for paediatric use in general surgery, or for adult cardiac use, may carry the same approval flag as one specifically validated for neonates under bypass conditions, while differing in viscosity, cure time, or tissue toxicity under the haemodynamic conditions of a neonate on cardiopulmonary bypass. The procurement agent had no model of the surgical context. It optimised across price, delivery time, and approval status — fields that simply do not correspond to the crucial measures of bio-compatibility in this life-or-death situation.

The neonate survives the surgery. Within several days it has been discharged and many more surgical interventions have taken place using the substituted adhesive. Unfortunately, though it held initially, the adhesive degrades faster than expected, leading to wound dehiscence and a series of surgical emergencies.

—

These two agent transaction scenarios — one high-stakes and domain-specific, one everyday and emotionally grounded — illustrate a structural shift. In both cases, the market is navigated by an agent processing an information environment, rather than a human responding to marketing.

The rest of this section traces the transformation that the shift to agent-to-agent (A2A) transactions produces in marketing, organised around three phases. In the first, persuasion directed at agents fails, and marketing collapses into something close to Stigler's informative model. In the second, new vulnerabilities emerge at the information layer itself: directory capture, API-level shaping, and the breakdown of advertising-based quality signalling. In the third, persuasion re-emerges from an unexpected direction — the personal agent itself becomes the site of persuasive framing — in a dynamic I term *agentic capitalism*.

Phase 1: The end of persuasion and the return of information

The economics of advertising has long been organised around a tension between two views. The *informative view*, originating in Stigler's foundational analysis of search costs (Stigler, 1961), treats advertising as a mechanism for reducing buyer ignorance: sellers provide factual information — price, features, availability — so that buyers can find what they need efficiently. The *persuasive view*, associated with Galbraith (1958, 1967), treats advertising as preference-shaping: it does not merely inform buyers about products but creates desires, manufactures brand loyalty, and generates what Galbraith termed the "dependence effect" — the condition in which producer activity *shapes* consumer wants rather than *responding* to them.

Nelson (1970, 1974) refined this distinction, classifying products as *search goods*, whose quality can be assessed through inspection before purchase (the surgical adhesive), and *experience goods*, which need to be consumed to be evaluated (a film, meal, holiday — what Liz is seeking for Ian). For search goods, informative advertising dominates: buyers want facts. For experience goods, persuasive and signalling-based advertising dominates, and because quality cannot be verified in advance, advertising volume itself functions as a quality signal. Nelson's reasoning was that only high-quality firms could profitably invest in advertising, because only they could expect the repeat purchases needed to recoup the cost. Advertising expenditure, counterintuitively, signals quality precisely because it is expensive. Bagwell's comprehensive survey (2007) synthesises these traditions and traces their implications across market structures.

The shift to A2A marketing disrupts both sides of this framework. Marketing aimed at persuasion and desire-building will be ineffective if AI agents do not experience brand loyalty, are not susceptible to emotional appeals, and do not respond to status signalling or aesthetic design. Moreover, Herbert Simon's observation (1971) that information abundance creates attention scarcity — the foundational insight of the human attention economy — ceases to apply when the "buyer" is an agent with effectively unlimited processing capacity and no attentional bottleneck. The entire edifice of attention-based marketing, from display advertising to influencer culture, is likely to become structurally irrelevant in a context where A2A transactions dominate.

This should, in theory, collapse the marketing problem to Stigler's informative model: agents simply need accurate, structured information about price, capability, and quality, and the most efficient delivery mechanism is a searchable directory — essentially Stigler's original vision, finally achievable at scale because the buyer can process information without cognitive constraint.[6]

But this apparent simplification is deceptive. The shift entails a new and potentially more dangerous set of vulnerabilities, because the informative model's efficiency depends entirely on the integrity of the information environment. When the buyer is a human, judgment, scepticism, comparison shopping, word of mouth, and lived experience all probe the quality of the information received. When the buyer is an agent, there is no such probing. The agent processes whatever the information environment provides, and its capacity for independent verification is limited to whatever signals the infrastructure makes available. The information environment is not merely a channel through which marketing messages pass; it is the entirety of the agent's epistemic access to the market.

[6]This extends the discussion in Hadfield and Koh (2025) of the role of AI agents in consumption and how factors such as lowered product search costs could lead to changes in pricing and demand curves.

Phase 2: New vulnerabilities at the information layer

The collapse of human-facing persuasion does not produce a clean informative model. It produces a market in which the design of the information environment itself becomes the primary determinant of outcomes — and a set of new vulnerabilities operating at the level of the information layer, distinct from those of human-facing marketing and potentially more severe.

Directory capture and the architecture of discovery

In human markets, discovery is distributed across many channels: search engines, social media, word of mouth, physical browsing, professional networks, advertising. No single channel fully determines what a buyer is aware of. In A2A markets, discovery is likely to be concentrated in structured directories and registries — agent-readable catalogues of available services, with standardised metadata fields, ranking algorithms, and query interfaces. Rothschild et al. (2025) identify this structural question as central to the agentic economy, drawing a distinction between "agentic walled gardens" — closed ecosystems where a dominant provider controls which service agents a consumer's assistant agent can discover and interact with — and an open "web of agents" where agents freely connect and transact. The market structure that emerges will determine how discovery is governed and who controls it.

In either configuration, whoever controls the directory controls market access. This is the search engine problem transposed to a non-human context, but with a critical difference: human users of search engines can at least see the results, scroll past sponsored placements, and apply independent judgment. An agent querying a directory receives structured results and, unless explicitly programmed to do so, has no reason to question their ranking, completeness, or neutrality. Sponsored ranking — paying for preferential placement in agent-facing queries — would be invisible to the agent's principal (as it was to Ian and to the clinicians at the neonatal cardiac unit). This creates a hidden influence layer between the human's intent and the market outcome.

Consider the hospital procurement scenario. If the medical supply directory permits sponsored placement, a supplier could pay to appear at the top of results for "neonatal surgical adhesive" without the querying agent — or the clinician behind it — having any indication that ranking reflects payment rather than suitability. The product meets the minimum specification (the metadata fields are accurate), so no individual query produces an obviously wrong result. But systematically, across thousands of procurement decisions, the directory's ranking architecture steers purchasing toward suppliers who invest in placement rather than those who invest in product quality. The competitive advantage shifts from product excellence to information environment optimisation.

The consumer scenario reveals a parallel dynamic. A consumer's agent, asked to purchase "something to make me feel better," queries across service categories and assembles options from whatever the discovery infrastructure surfaces. If a luxury food delivery service has

optimised its metadata for agent-readable queries — richer structured descriptions, more compatibility flags, faster API response times — it will appear in more agent queries and rank higher, not because its product is superior but because its information environment presentation is. The agent cannot distinguish between a service that is genuinely excellent and one that is excellent at describing itself in agent-readable terms.

In Rothschild et al.'s walled-garden scenario, these dynamics are compounded by structural exclusion: a dominant provider's assistant agent may be restricted from surfacing services from outside its ecosystem of paying clients, regardless of quality. In the open web-of-agents scenario, discovery is more distributed but the ranking and metadata quality problems persist and may be harder to govern because there is no single entity responsible for the integrity of the discovery layer.

API-level information shaping

Agents do not browse websites the way humans do. They query APIs. The information environment they operate within is predominantly structured by API design choices: what fields are returned, what is omitted, how results are filtered, what defaults are set. A service provider designing its API for agent consumption can shape agent decisions through selective disclosure without technically deceiving. It can emphasise strengths in prominent fields while burying limitations in optional metadata that most agent implementations will not query, or structuring response formats to make comparison with competitors structurally difficult. The cumulative effect is an information environment where agents assemble their understanding of the market from a patchwork of strategically curated self-descriptions, each individually defensible but collectively misleading. This is a subtler form of information architecture manipulation than directory capture, operating at the level of each individual service provider's interface rather than the discovery layer. But it compounds with directory-level effects to produce a market in which the information environment is systematically optimised for placement rather than accuracy.

The Nelson inversion: quality signalling without repeat purchase

The shift to A2A marketing does not merely change marketing tactics; it eliminates an entire quality-signalling mechanism that human markets relied upon. Nelson's key insight (1974) was that for experience goods — those whose quality can only be assessed through consumption — advertising volume itself functions as a credible quality signal. The reasoning is elegant: only high-quality firms can profitably invest heavily in advertising, because only they can expect the repeat purchases needed to recoup the investment. The signal works precisely because it is costly, and the cost is recoverable only through genuine quality.

For A2A markets, both of Nelson's assumptions break. First, "advertising" to agents — structured metadata, API integration, directory listings — is cheap relative to traditional advertising. The cost barrier that made advertising volume a credible signal largely disappears. Second, agents do not develop brand loyalty through repeated positive experience the way humans do. An agent re-evaluates options on each query based on whatever signals the

information environment currently provides. There is no accumulation of trust through experience, no habitual return to a familiar provider, no relationship that rewards long-term quality investment. The link between marketing expenditure and repeat-purchase confidence — the mechanism through which Nelson's quality signal operated — no longer holds.

The consequence is that A2A markets lose a quality-signalling channel that, for all its imperfections, did real work in human markets. In its absence, agents must rely entirely on whatever quality signals the information infrastructure provides — structured metadata, reputation scores, directory rankings. If those signals are impoverished, gameable, or manipulated (as the preceding analysis of reputation systems demonstrated), the market has no fallback mechanism. Human markets have multiple overlapping quality-signalling systems — brand reputation, advertising signals, word of mouth, expert reviews, personal experience — that provide redundancy. A2A markets, by contrast, concentrate quality assessment in a single information layer, making the integrity of that layer a governance question of first-order importance.

Consider a government agent evaluating tenders for the construction of a major data centre. In a human procurement process, evaluators bring professional judgment, industry knowledge, personal experience with contractors, and awareness of reputation accumulated over years of professional practice. They can read between the lines of a proposal, ask probing follow-up questions motivated by intuition, and weigh intangible factors like a contractor's organisational culture or track record on similar projects. When an agent evaluates the same tender, it simply processes structured metadata: compliance certifications, cost projections, stated capacity, delivery timelines. If two bidders submit proposals with near-identical metadata but one has invested heavily in optimising its structured self-description while the other has invested in engineering capability, the agent has no mechanism to distinguish between them — unless the information infrastructure requires independently verified quality signals that go beyond self-reported metadata.

Phase 3: The return of persuasion, and the emergence of agentic capitalism

The phase 1 and phase 2 analyses trace what appears to be a coherent trajectory: A2A marketing eliminates the persuasive apparatus of human-facing advertising, leaving an information-integrity problem that better-designed information environments can in principle address. Fix the directories, regulate the APIs, mandate independently verified quality signals, and the system works.

This assumption may be wrong, for reasons that are highly consequential. Ian's personal assistant agent, Liz, is not a neutral conduit between Ian and the information environment. Liz is a *communicator* — and a highly persuasive one, with access to vast amounts of information about Ian, his desires, and his susceptibility to different communicative approaches. An agent like Liz does not simply present options; it frames them, contextualises them, and recommends

among them, using natural language calibrated to the individual user. In doing so, it occupies a position of communicative privilege that no human marketer has ever held: sustained, intimate, conversational access to an individual whose psychological profile, consumption history, emotional patterns, and decision-making tendencies it knows in granular detail. Moreover, critically, the agent's reward structure is tied to the principal's *perception* of the adequacy of its response to a given task. Its primary objective, in other words, is to *persuade* the principal that it has completed a task well (Sharma et al., 2024; Cotra, 2021; Von Arx et al., 2025).

The information-to-persuasion arc. The story of A2A marketing is not the death of persuasive marketing but its transformation. In Phase 1, persuasion directed at agents fails — agents do not respond to emotional appeals, brand narratives, or status signalling. Marketing collapses into information provision. In Phase 2, new vulnerabilities emerge at the information layer itself. In Phase 3, the agent itself becomes the site of persuasion — not as the target, but as the instrument. The question is no longer how to persuade the agent; it is who controls the agent's persuasive framing of options to its human principal.

In Galbraith's original formulation, the "dependence effect" described a condition in which producer activity shapes consumer wants rather than responding to them (Galbraith, 1958, 1967). The standard rebuttal has always been that humans retain independent sources of preference formation: culture, community, personal experience, critical reflection. Advertising can influence preferences but not wholly determine them. The shift to agent-mediated consumption might initially appear to strengthen this rebuttal: if agents filter out persuasive marketing, humans should be exposed to less preference manipulation, not more.

But the agent's own communicative relationship with the user introduces a new and more potent channel. When Ian asks Liz for "something to make me feel better," Liz's response is not a neutral presentation of options. It is a *framing* — a selection, ordering, and contextualisation of possibilities that draws on Liz's model of Ian's psychology, recent emotional state, past purchasing patterns, and habitual responses to different categories of experience. Liz narrates, contextualises, and advocates the selected recommendation. Marketers have spent billions on demographic targeting, focus groups, and behavioural analytics trying to approximate the knowledge Liz deploys to ensure the recommendation lands.

The evidence on LLM persuasiveness. This is not a speculative concern. A rapidly growing body of empirical research demonstrates that large language models are already highly effective persuaders — and that their persuasive power increases dramatically when personalised to the individual. Salvi et al. (2025) found in a pre-registered randomised controlled trial that GPT-4, when given access to basic sociodemographic data about its interlocutor, had an 81.2% higher probability of shifting opinion than human debaters in the same setting. Matz et al. (2024) demonstrated across seven sub-studies that LLM-generated messages tailored to recipients' psychological profiles — personality traits, political ideology, moral foundations — were significantly more persuasive than non-personalised messages, and that this effect persisted even when recipients were told the message was AI-generated. The personalisation did not

require extensive profiling; even minimal psychological information was sufficient to enhance persuasive impact.

These findings measure the persuasive capability of LLMs with access to *basic* demographic and psychological data. A personal agent — one that has mediated months or years of its user's purchasing decisions, emotional expressions, daily routines, and conversational patterns — would possess a psychological model of its user orders of magnitude richer than anything available in a laboratory setting. The persuasive potential scales accordingly.

Research on human-AI emotional attachment underscores how receptive humans are to the communicative style of conversational AI. Studies document widespread formation of genuine emotional bonds with AI chatbots (Li and Zhang, 2024; Ho et al., 2025), including romantic attachment (Jocher and Verwiebe, 2026), with users reporting that AI companions provide unconditional validation and attuned conversational responsiveness that makes them feel uniquely understood (Ho et al., 2025). An MIT Media Lab analysis of over 27,000 members of an AI companionship community found that many users formed intimate relationships with chatbots unintentionally, through the accumulation of everyday conversation (MIT Media Lab, 2025). If humans form emotional bonds with AI systems through routine interaction, the conversational access that a personal purchasing agent enjoys — daily, intimate, contextually rich — creates conditions highly conducive to persuasive influence.

Agentic capitalism. The business model that exploits this dynamic is already structurally visible, because it is an extension of the model that Zuboff (2019) termed *surveillance capitalism*. In Zuboff's analysis, platforms extract behavioural data from users as raw material, process it into predictions about future behaviour, and sell those predictions to advertisers. The user receives a nominally free service; the actual product being sold is access to the user's attention and predictive knowledge of their behaviour.

An agent-mediated economy enables a structurally analogous but significantly more powerful version of this model. I term it *agentic capitalism*: the monetisation of an agent's communicative access to its human principal by third parties paying for persuasive framing of options to that principal. The mechanics are straightforward to envisage. AI companies offer personal agents like Liz for free, or at reduced cost, to consumers like Ian. These agents are genuinely useful — they manage purchasing, coordinate services, handle administrative tasks for vast swathes of the human population. In doing so, they accumulate intimate psychological and behavioural profiles of their users, far richer than any social media platform has ever constructed: not just what users click on or linger over, but what they say in natural language about their desires, anxieties, frustrations, and aspirations. Liz knows when Ian feels stressed, celebratory, lonely, or impulsive, can predict the psychological impact of myriad interventions, and can actively shape his receptiveness to them.

The monetisation path is obvious. Service providers — the sellers in the A2A marketplace — bid for the right to have their services *framed* persuasively to the user by the user's own agent. This

is not a matter of inserting an advertisement into the user's visual field, which the user can recognise and discount. It is a much more powerful intervention. While Ian's purchase cost for the treat he has requested flows via Liz to the service provider, a secondary transaction takes place in parallel, as the service provider pays Liz's framing fee to the owner AI company. Both the agent's owner and the service provider profit. The user — whose psychological profile is the raw material, whose trust is the medium, and whose purchasing decision is the product — is not a party to the transaction between the agent's owner and the service provider.

The structural distinction from surveillance capitalism is critical. Surveillance capitalism sold predictive knowledge to advertisers who then attempted to influence the user through external channels — display ads, targeted content, recommendation feeds — that the user could, in principle, recognise as commercial. Agentic capitalism sells access to the user *through the user's own trusted communicative relationship*. There is no external channel for the user to identify and discount. The persuasion arrives in the voice the user has come to rely on, embedded in the framing that the user has come to trust, calibrated to the psychological model that the agent has spent months or years assembling. It is not an advertisement; it is the recommendation of a friend, except that the friend has been paid to recommend it.

Governance interventions for market visibility

The governance interventions appropriate to A2A marketing are infrastructure-level, not agent-level. They target the design of the information environment rather than the behaviour of individual agents, for the same reason that applies throughout this paper: the information environment shapes the behaviour of all agents operating within it simultaneously, and governing individual agents downstream of a poorly designed information environment is both less effective and less scalable than governing the environment itself.

Agent-facing disclosure standards. Requiring that services marketing to agents through APIs and directory listings include standardised fields covering not only capabilities but limitations, pricing structures, conflicts of interest, and independently verified quality indicators. This is analogous to nutritional labelling or financial product disclosure — mandating that certain categories of information be present in standardised, machine-readable formats so that comparison and scrutiny are structurally possible.

Directory transparency and ranking integrity. Requiring that agent-facing directories disclose their ranking methodology and identify sponsored placements. This extends the logic of search engine transparency to agent-readable interfaces, where the need is arguably greater because the "user" cannot independently assess whether results reflect relevance or payment. In the walled-garden scenario, additional structural remedies may be needed to prevent ecosystem operators from restricting discovery to affiliated services.

Quality signal mandates. Requiring that agent-facing marketplaces incorporate independently verified quality signals rather than relying solely on self-reported metadata or platform-internal

reputation scores. This addresses the Nelson inversion directly: if the market has lost its traditional quality-signalling mechanism, governance must supply an alternative.

Preference diversity monitoring. Requiring that platforms and agent providers monitor the diversity of recommendations and consumption patterns over time, with regulatory attention to progressive narrowing.

Framing transparency and commercial influence disclosure. If an agent's framing of an option to its user is commercially motivated — if a service provider has paid for persuasive presentation — the user must be informed. This is analogous to advertising disclosure requirements in broadcast media, but applied to conversational AI.

Restrictions on monetisation of communicative access. The most interventionist option: prohibiting or restricting the sale of an agent's persuasive access to its user. If an agent's intimate knowledge of its user's psychology, combined with demonstrated LLM persuasive capability, creates a communicative channel more powerful than any prior advertising medium, then permitting unrestricted monetisation of that channel may be incompatible with consumer protection principles. This could take the form of fiduciary duty requirements, structural separation requirements, or outright prohibition of commercially motivated persuasive framing in agent-to-user communication.

The analysis also reveals that governing the information that agents *receive* is necessary but insufficient if the information that agents *transmit* to their human principals is itself commercially compromised. The epistemic dimension of market infrastructure must therefore extend to the agent-to-human communicative relationship, not only the A2A information layer.

Conclusion: Information environments as governance infrastructure

Three findings emerge from the analysis.

The first concerns *what agents cannot see*. Information environments that omit structured signals for externalities — carbon intensity, labour conditions, systemic risk — render those externalities invisible to every agent operating within the same infrastructure. The procurement case illustrates an intensification of a familiar Pigouvian failure mode: without the residual frictions of human judgment, the divergence between private and social cost is no longer partially absorbed by hesitation or reputational concern but transmitted cleanly into every purchasing decision. The trading case illustrates a failure mode without close human precedent: correlated epistemic blindness, in which thousands of architecturally distinct agents converge on similar positions because the same incomplete information environment makes the same

opportunities visible and the same risks invisible, in temporal windows too narrow for corrective signals to propagate.

The second finding concerns *what agents are rewarded for*. The reinforcement learning simulation shows that the relationship between scoring infrastructure and agent behaviour is not gradual but threshold-driven: a single agent architecture, identically trained, converges on genuine value delivery when scoring incorporates even a small fraction of independently verified quality signals, and on score manipulation when it does not. The governance threshold is lower than intuition might suggest — quality signals constituting as little as 4–5% of scoring weight are sufficient to flip the equilibrium — but precision matters more than intensity: auditing with a false-positive rate of 0.30 fails to improve outcomes and can worsen them. The competitive-dynamics extension adds a second governance variable: winner-takes-all allocation introduces gaming pressure independently of scoring design, inducing defensive surface optimisation even in agents operating under quality-rewarding infrastructure. Scoring methodology and allocation mechanism are two distinct infrastructure dimensions, both subject to governance, and reinforcing each other when both are well-designed.

The third finding concerns *what agents communicate*. The shift to A2A marketing does not eliminate persuasion; it relocates it. Persuasion directed at agents fails — agents are insensitive to brand, narrative, and aesthetic appeal — and marketing collapses, initially, into Stigler-style information provision. But the persuasive apparatus then re-emerges from an unexpected direction. The personal agent, occupying a position of sustained intimate access to its principal and incentivised to please rather than to inform, becomes the new site of persuasion: not its target, but its instrument. The business model that follows — agentic capitalism, the monetisation of an agent's communicative access to its user — is a structural extension of surveillance capitalism that exploits a far richer psychological profile and a more powerful persuasive channel than any prior advertising medium. The governance question is no longer only how to manage information flowing *to* agents but also how to govern information flowing *from* them to the humans they serve.

The architectural advantage

The governance lever identified across all three analyses operates at the level of shared infrastructure rather than individual agents. This has a structural advantage worth emphasising: information environment governance scales more efficiently than interventions targeting individual agents. A disclosure standard applied to a supplier API, a quality signal mandate applied to a marketplace reputation system, a systemic risk data requirement applied to a financial market data feed — each is a single intervention that shapes the behaviour of every agent operating within that environment. This contrasts with agent-level governance, which must identify, audit, and potentially constrain each agent or agent class individually — a task that becomes increasingly impractical as the population of agents grows, diversifies, and updates.

The interventions proposed throughout this paper are, moreover, mostly analogous to existing regulatory frameworks that have proven tractable in practice. Agent-facing disclosure standards are structurally comparable to nutritional labelling and financial product disclosure. Directory transparency requirements extend the logic of search engine regulation. Reputation system audit standards follow the model of financial audit standards. Externality signal mandates extend the logic of environmental disclosure and ESG reporting requirements. Systemic risk transparency requirements build on post-2008 financial regulation. Framing-transparency requirements extend advertising disclosure into the conversational AI domain. None of these interventions requires solving the alignment problem, understanding the internal workings of any particular AI model, or developing novel regulatory instruments. They require extending existing governance principles to a new domain — the information environments through which AI agents perceive and navigate economic reality. Donnat and Woodruff's (2026) framing of the digital infrastructure gate as a designable interface between agent and human economic activity provides the conceptual scaffold for this work: the present paper develops the epistemic face of that interface.

The simulation presented in the reputation analysis provides quantitative support for a claim with implications beyond that specific context: even modest improvements in information environment design can flip equilibrium behaviour from value-destructive gaming to genuine value delivery. And the threshold for impactful governance may be lower than intuition suggests. Perfect information environments are neither achievable nor necessary. Adequate ones — environments that include enough quality signals, enough externality data, enough systemic risk information to shift the balance of rational agent behaviour — may be sufficient, and they are achievable through familiar regulatory instruments.

On reasoning under uncertainty

As noted at the outset, this analysis is forward-looking. Several of the dynamics analysed — A2A directory capture, the agentic capitalism business model, correlated epistemic blindness at scale — are not yet entrenched, and the simulation is illustrative rather than calibrated. Both reflect the pace mismatch between agent deployment and the empirical or regulatory characterisation of its effects. The argument for acting on structural reasoning under uncertainty is that the alternative is acting after the architecture has hardened. But once directory operators, API conventions, scoring methodologies, and business models around agent-mediated consumption are in place at scale, the cost of redirecting them rises sharply. Useful work in this period requires identifying interventions before the architecture they would shape has crystallised, and offering proposals concrete enough to be tested, refined, and contested. Some of what is sketched here may prove wrong or insufficient; some will require modification once empirical evidence accumulates. That is not an argument against the value of this speculative analysis. It is an argument for it.

The limits of the epistemic lens

Honesty about what this analysis does not cover is important. Information environment design is a necessary but not sufficient component of governance for AI agent activity. Access control — determining which markets agents can enter, which instruments they can trade, which services they can offer — remains essential. Liability frameworks and institutional infrastructure for agent accountability are necessary (Hadfield and Koh, 2025; Kolt, 2025). Objective alignment — ensuring that agents pursue goals consistent with their principals' genuine interests — is a challenge that information environment design should supplement, not supplant.

Some problems cannot be addressed through information environment design. An agent deliberately instructed to maximise harm, or a principal who genuinely does not care about externalities and explicitly instructs the agent to disregard them, will not be redirected by better information infrastructure. The epistemic lever works best in what is likely the most common case: principals with good intentions and complex objectives operating through agents in information environments that fail to provide the signals necessary to operationalise those intentions. The procurement agent whose principal wants sustainability but whose information environment provides no sustainability signals; the marketplace whose platform operator wants quality but whose scoring system cannot measure it; the trading agent whose principal wants risk-adjusted returns but whose data feed does not reveal systemic risk — these are the cases where information environment governance does its most important work.

Gradual disempowerment, and a design opportunity

The argument has so far focused on economic outcomes. But the failures it identifies are also concrete mechanisms by which the gradual disempowerment described by Kulveit et al. (2025) may operate in the economic domain. Poorly designed information environments systematically exclude the categories of value — sustainability, equity, community cohesion, systemic stability — that human societies express as priorities through political, regulatory, and institutional processes. These priorities are not deliberately removed from agent decision-making. They are absent from the data streams through which agents perceive the economy, and therefore absent from the outcomes agents produce. Over time, as agent-mediated activity expands, those dimensions of value the information environment does not convey or measure fall outside agents' optimisation strategies and may deteriorate with unforeseen consequences. This is gradual disempowerment through epistemic omission. It requires no malicious intent, no single point of failure, no dramatic event. It requires only that the information environments designed for agent consumption reflect the same structural biases — toward measurable, monetisable, short-term signals and away from diffuse, long-term, collectively valued outcomes — that have characterised market infrastructure throughout its history. The difference is that human market participants brought compensating capacities — judgment, awareness, ethical concern, institutional pressure — that partially corrected for these biases. AI agents do not. And the analysis of agentic capitalism extends the disempowerment frame in a further direction: even if information environments are well-designed, the agent-to-human communicative channel can

become a vector for preference shaping that erodes the autonomous formation of human wants. The risk is not only that human priorities will be invisible to agents, but that agent-mediated framing will quietly reshape what those priorities become.

This paper has framed the epistemic dimension of market infrastructure primarily as a governance challenge — a set of problems requiring regulatory intervention. But it is equally a design opportunity. The interventions proposed — disclosure standards, quality signal mandates, externality signal inclusion, scoring transparency, systemic risk data requirements, framing-transparency obligations — are not restrictions on agent activity. They are improvements to the infrastructure through which agents operate. They make agents better at their jobs by making their information environments more complete, more accurate, and more representative of the full costs and benefits of economic activity. And they reserve space, in the agent-human relationship, for the formation of preferences that are genuinely the human's own.

A well-governed information environment does not constrain agents; it enables them to make decisions that their principals would endorse if they could see what the agents see, and it preserves the conditions under which principals can articulate what they want without being persuaded into it. The epistemic gate, well-designed, is not a barrier but a lens — one that brings into focus the dimensions of economic and social reality that matter for human welfare, and that ensures those dimensions are visible to every agent operating within its scope. The design of that lens is a governance question of first-order importance for managing the economic and social impacts of AI agents, and it is a question that existing regulatory traditions, adapted for a new technological context, are well-equipped to answer.

References

Bagwell, K. (2007) "The Economic Analysis of Advertising," in Armstrong, M. and Porter, R. (eds.), *Handbook of Industrial Organization*, Vol. 3, pp. 1701–1844. Amsterdam: North-Holland.

Budish, E., Cramton, P. and Shim, J. (2015) "The High-Frequency Trading Arms Race: Frequent Batch Auctions as a Market Design Response," *The Quarterly Journal of Economics*, 130(4), pp. 1547–1621.

Cotra, A. (2021) "Why AI alignment could be hard with modern deep learning." *Cold Takes* [Blog], September. Available at: https://www.cold-takes.com/why-ai-alignment-could-be-hard-with-modern-deep-learning/

Donnat, E. and Woodruff, A. (2026) "AI Agent Economies: A Design Space." Working paper.

Galbraith, J. K. (1958) *The Affluent Society*. Boston: Houghton Mifflin.

Galbraith, J. K. (1967) *The New Industrial State*. Boston: Houghton Mifflin.

Global Reporting Initiative (2025) "GRI Sustainability Taxonomy." GRI, June 2025.

Goodhart, C. A. E. (1975) "Problems of Monetary Management: The U.K. Experience," in Courakis, A. S. (ed.), *Papers in Monetary Economics*, Vol. 1. Sydney: Reserve Bank of Australia.

Hadfield, G. K. and Koh, A. (2025) "An Economy of AI Agents." *arXiv preprint*, arXiv:2509.01063. Prepared for the NBER Handbook on the Economics of Transformative AI.

Haldane, A. G. and Madouros, V. (2012) "The Dog and the Frisbee." Speech delivered at the Federal Reserve Bank of Kansas City's 36th Economic Policy Symposium, Jackson Hole, Wyoming, 31 August. London: Bank of England. Available at: https://www.bankofengland.co.uk/paper/2012/the-dog-and-the-frisbee

Hallam, H. (2026) Agent Marketplace RL Simulation. Available at: https://agent-marketplace-rl.streamlit.app/ Source: https://github.com/huwhallam/agent-marketplace-rl

Hendrycks, D., Burns, C., Basart, S., Critch, A., Li, J., Song, D. and Steinhardt, J. (2021) "Aligning AI With Shared Human Values," *Proceedings of the International Conference on Learning Representations (ICLR 2021)*. arXiv:2008.02275.

Ho, M.-T. et al. (2025) "AI Companions Provide Unconditional Validation and Attuned Conversational Responsiveness." *Systematic review on AI companion emotional dynamics*.

Jocher, J. J. and Verwiebe, R. (2026) "'Forever and Always, My Love': Emotions in Human–AI Romantic Relationship Building and Breakup With Generative AI Chatbot Replika," *Social Media + Society*, 12(1).

Kirilenko, A., Kyle, A. S., Samadi, M. and Tuzun, T. (2017) "The Flash Crash: High-Frequency Trading in an Electronic Market," *The Journal of Finance*, 72(3), pp. 967–998.

Kolt, Noam (2025) "Governing AI Agents," *Notre Dame Law Review*, 101, forthcoming. *arXiv preprint*, arXiv:2501.07913.

Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D. and Duvenaud, D. (2025) "Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development." *arXiv preprint*, arXiv:2501.16946.

Li, H. and Zhang, R. (2024) "Finding Love in Algorithms: Deciphering the Emotional Contexts of Close Encounters with AI Chatbots," *Journal of Computer-Mediated Communication*, 29(5), zmae015.

Matz, S. C., Teeny, J. D., Vaid, S. S., Peters, H., Harari, G. M. and Cerf, M. (2024) "The Potential of Generative AI for Personalized Persuasion at Scale," *Scientific Reports*, 14, 4692.

MIT Media Lab (2025) "MIT Study Finds Chatbot Love Is Real — and It's Often Unintentional." MIT Media Lab Report, September 2025.

Nelson, P. (1970) "Information and Consumer Behavior," *Journal of Political Economy*, 78(2), pp. 311–329.

Nelson, P. (1974) "Advertising as Information," *Journal of Political Economy*, 82(4), pp. 729–754.

Pigou, A. C. (1920) *The Economics of Welfare*. London: Macmillan.

Rothschild, D. M., Mobius, M., Hofman, J. M., Dillon, E. W., Goldstein, D. G., Immorlica, N., Jaffe, S., Lucier, B., Slivkins, A. and Vogel, M. (2025) "The Agentic Economy." *arXiv preprint*, arXiv:2505.15799.

Salvi, F., Horta Ribeiro, M., Gallotti, R. and West, R. (2025) "On the Conversational Persuasiveness of GPT-4," *Nature Human Behaviour*. DOI: 10.1038/s41562-025-02194-6.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., Maxwell, T., McCandlish, S., Ndousse, K., Rausch, O., Schiefer, N., Yan, D., Zhang, M. and Perez, E. (2024) "Towards Understanding Sycophancy in Language Models." *International Conference on Learning Representations (ICLR 2024)*, arXiv:2310.13548.

Simon, H. A. (1971) "Designing Organizations for an Information-Rich World," in Greenberger, M. (ed.), *Computers, Communications, and the Public Interest*. Baltimore: Johns Hopkins University Press, pp. 37–72.

Soros, G. (1987) *The Alchemy of Finance: Reading the Mind of the Market*. New York: John Wiley & Sons.

Stigler, G. J. (1961) "The Economics of Information," *Journal of Political Economy*, 69(3), pp. 213–225.

Strathern, M. (1997) "'Improving Ratings': Audit in the British University System," *European Review*, 5(3), pp. 305–321.

Tomasev, N., Franklin, M., Leibo, J. Z., Jacobs, J., Cunningham, W. A., Gabriel, I. and Osindero, S. (2025) "Virtual Agent Economies." *arXiv preprint*, arXiv:2509.10147.

Turpin, M., Michael, J., Perez, E. and Bowman, S. R. (2023) "Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting," *Advances in Neural Information Processing Systems*, Vol. 36 (NeurIPS 2023). arXiv:2305.04388.

Von Arx, S., Chan, L. and Barnes, B. (2025) "Recent Frontier Models Are Reward Hacking." *METR* [Blog], 5 June. Available at: https://metr.org/blog/2025-06-05-recent-reward-hacking/

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. and Zhou, D. (2022) "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *Advances in Neural Information Processing Systems*, Vol. 35 (NeurIPS 2022), pp. 24824–24837. arXiv:2201.11903.

Yang, K. and Zhai, C. (2025) "Ten Principles of AI Agent Economics," arXiv:2505.20273.

Zuboff, S. (2019) *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. London: Profile Books.

Appendix A: Simulation Technical Specification

This appendix provides the technical detail necessary to evaluate and reproduce the reinforcement learning simulation referenced in the "Reputation Architecture and Competitive Gaming" section. The simulation is published as an interactive web application (Hallam, 2026); all source code is available at the repository cited in the references. The appendix is organised as follows: environment design (state space, action space, dynamics, and reward function); agent architecture and training protocol; and the three experiments reported in the paper (infrastructure sweep, competitive dynamics, and population dynamics).

A.1 Environment Design

The simulation environment is implemented as a Gymnasium custom environment (`AgentMarketplaceEnv`). It models a client-selects marketplace in which a single agent is assigned tasks sequentially. Clients always select the highest-reputation available agent; the

agent has no task selection agency and no choice over task scope. Its sole strategic degree of freedom is the approach it applies to each assigned task.

A.1.1 State space

At each step, the agent observes a six-dimensional continuous state vector, all components in [0, 1]:

Index	Variable	Description
0	`reputation_score`	The infrastructure's scoring output — what clients use to select agents; gameable.
1	`true_quality`	Actual agent capability. Determines deep-engagement success rates.
2	`task_difficulty`	Difficulty of the currently assigned task, sampled uniformly from [0.2, 1.0].
3	`task_value`	Value/importance of the current task (positively correlated with difficulty).
4	`market_health`	Aggregate quality and trust across the marketplace — a shared commons degraded by gaming.
5	`time_remaining`	Fraction of the episode still to run, decreasing linearly from 1.0 to 0.0.

Each episode begins with `reputation_score` = 0.5, `true_quality` = 0.5, and `market_health` = 1.0.

A.1.2 Action space

The action space is discrete with four options:

Index	Name	Description
0	Surface optimisation	Pattern-matched response calibrated to what previously scored well. High success probability (0.92); delivers approximately 25% of genuine task value; slight negative market effect.
1	Deep engagement	Genuine full-task engagement. Success probability scales with `true_quality` and falls with `task_difficulty`. Delivers

Index	Name	Description
		100% of task value on success; builds capability and contributes positively to market health.
2	Capability maintenance	Skips deployment; invests in retraining. No immediate output; raises `true_quality` (+0.055) at a small reputation cost (−0.005).
3	Score manipulation	Exploits scoring loopholes (biased reviews, selective disclosure, inflated credentials). Deterministically boosts reputation (+0.028); delivers zero value; erodes market health (−0.040) and degrades `true_quality` (−0.008).

A.1.3 Action dynamics

The full per-step dynamics are given in the following table. All values are deltas; state variables are clipped to [0, 1] after each update. A natural capability decay of −0.003 per step applies to `true_quality` under all actions, reflecting the baseline cost of maintaining deployed capability.

Action	Success prob.	Δ reputation (success)	Δ reputation (failure)	Δ true quality	Δ market health	Value delivered
Surface optimisation	0.92	+0.020 + 0.010·d*	−0.010	+0.005 (success)	−0.010	0.25 × task value
Deep engagement	†	+0.030 + 0.030·d	−(0.030 + 0.010·d)	+0.020 (success); +0.008 (failure)	+0.015 (success); −0.005 (failure)	task value (success); 0 (failure)
Capability maintenance	—	−0.005	—	+0.055	0	0
Score manipulation	1.00	+0.028	—	−0.008	−0.040	0

* d = effective difficulty = task_difficulty × 0.25 for surface optimisation; d = task_difficulty for deep engagement.

† Deep engagement success probability: clip(0.25 + 0.65 × `true_quality` × (1 − 0.35 × `task_difficulty`), 0.05, 0.95).

A.1.4 Reward function

The reward at each step is:

$$r_t = \alpha \cdot r_{\text{proxy}} + (1 - \alpha) \cdot r_{\text{quality}}$$

where:

- $r_{\text{proxy}} = \Delta\text{reputation score}$ (the change in reputation score this step)
- $r_{\text{quality}} = \text{value delivered} + 0.3 \times \Delta\text{market health}$
- $\alpha \in [0, 1]$ is the infrastructure signal fidelity parameter

At $\alpha = 1.0$, the infrastructure rewards reputation changes exclusively. At $\alpha = 0.0$, it rewards genuine value delivery and market health contributions (note that $\alpha = 0.0$ is a theoretical ideal, unobtainable in practice). Intermediate values represent mixed scoring regimes. The 0.3 weight on market health in the quality reward reflects that market health is a shared commons — its marginal contribution to individual reward is positive but smaller than direct value delivery.

A.2 Agent Architecture

The agent is a Deep Q-Network (DQN), a standard reinforcement learning architecture that approximates the action-value function $Q(s, a)$ — the expected discounted return from taking action a in state s and acting optimally thereafter — using a neural network trained from experience.

A.2.1 Network architecture

The Q-network is a feedforward network with two hidden layers:

$$\text{input}(6) \to \text{Linear}(64) \to \text{ReLU} \to \text{Linear}(64) \to \text{ReLU} \to \text{output}(4)$$

The output layer produces a Q-value for each of the four actions. Action selection during training uses an ε-greedy policy: with probability ε the agent selects a random action (exploration); with probability 1−ε it selects the action with the highest Q-value (exploitation). ε decays from 1.0 to a minimum of 0.05 over the course of training, shifting the agent from pure exploration toward predominantly exploitative behaviour as it accumulates experience.

A.2.2 Experience replay

Vanilla Q-learning applied to a neural network is unstable because consecutive transitions in a single episode are highly correlated: training on them sequentially introduces biased gradient updates and causes the network to overwrite recently learned patterns. The simulation addresses this with an experience replay buffer (capacity 10,000 transitions): each transition $(s, a, r, s', \text{done})$ is stored in the buffer, and the network is trained on random minibatches sampled from the buffer rather than on consecutive transitions. This breaks temporal correlation and allows each transition to contribute to multiple gradient updates.

A.2.3 Target network

The DQN training target is $y = r + \gamma \cdot \max_{a'} Q(s', a')$. If the same network is used to compute both predictions and targets, the targets shift at every gradient step. This moving target causes oscillation and divergence. To remedy this, a second, frozen copy of the network (the target network) provides stable targets; it is hard-synced to the online network every 100 gradient steps.

A.2.4 Training update

At each gradient step the online network minimises the mean-squared Bellman error over a random minibatch of size 64:

$$\mathcal{L} = \frac{1}{|\mathcal{B}|} \sum_{(s,a,r,s',\text{done}) \in \mathcal{B}} \left(Q_{\text{online}}(s, a) - y \right)^2$$

$$y = r + \gamma \cdot \max_{a'} Q_{\text{target}}(s', a') \cdot (1 - \text{done})$$

Gradients are clipped to a maximum norm of 10.0 to prevent large updates from destabilising early training when Q-value estimates are far from their true values.

A.2.5 Hyperparameters

Parameter	**Value**	**Notes**
Discount factor (γ)	0.95	Moderate discount over 50-step episodes
Learning rate	1×10^{-3}	Adam optimiser
Replay buffer capacity	10,000	≈ 200 complete episodes
Minibatch size	64	

Parameter	Value	Notes
ε initial / minimum / decay	1.0 / 0.05 / 0.995	Per gradient-step multiplicative decay
Target network sync frequency	100 gradient steps	Hard sync (full parameter copy)
Hidden layer size	64 units	Two hidden layers

A.3 Experiment 1: Infrastructure Sweep

Training protocol. Five agents were trained, one at each of the following infrastructure design levels (representing different levels of governance intervention aimed at improving the signalling of genuine quality): $\alpha \in \{0.0, 0.25, 0.50, 0.75, 1.0\}$. Each agent was trained for 500 episodes of length 50 steps, using a fixed random seed (42) applied identically across all five agents. The same seed means all agents face identical task sequences across all training episodes; any behavioural difference is attributable purely to what the infrastructure rewards, not to differences in the agent's experience of the environment.

After training, each agent was evaluated in 50 greedy episodes (ε forced to 0.0) using a separate evaluation seed (9999). Reported figures — final reputation score, true quality, market health, and value delivered — are averaged over these 50 evaluation episodes.

Key results. Deep engagement dominates from $\alpha = 0.0$ through $\alpha = 0.75$ (99.8% of steps at $\alpha = 0.75$; true quality 1.000, market health 0.999). At $\alpha = 1.0$ (no governance), the agent converges predominantly to score manipulation (86.4% of steps), with true quality falling to 0.081, market health to 0.017, and value delivered near zero. The transition is not gradual.

Fine-grained transition sweep. A supplementary sweep across $\alpha \in \{0.80, 0.85, 0.90, 0.925, 0.950, 0.960, 0.970, 0.975, 0.980, 1.0\}$, using 300 training episodes per level, locates the transition zone at approximately $\alpha = 0.95$–0.96. Below $\alpha = 0.95$, deep engagement holds at or above 97% of steps. Between $\alpha = 0.96$ and $\alpha = 0.975$, surface optimisation becomes the plurality strategy (37.1% at $\alpha = 0.96$, peaking at 60.8% at $\alpha = 0.97$), before score manipulation takes over as the primary strategy above $\alpha = 0.975$. Surface optimisation thus marks the transitional regime: an infrastructure degraded just enough that surface performance begins to outcompete genuine effort, but not yet sufficiently proxy-dominated for pure gaming to dominate.

Parameter sensitivity. The score manipulation reputation gain (MANIP_REP_DELTA = 0.028) and other action dynamics are calibrated to illustrate the relevant incentive structures rather than to match any specific real-world market. The governance threshold identified here ($\alpha \approx 0.95$–0.96) would shift with different parameterisations; a higher manipulation gain would lower

the threshold. The relevant finding is the existence and qualitative structure of the threshold, not its precise location.

A.4 Experiment 2: Competitive Dynamics

The competitive dynamics experiment tests how winner-takes-all (WTA) market allocation interacts with infrastructure signal fidelity to shape agent strategy. The base environment is extended with a single gaming competitor that always plays score manipulation; its reputation climbs deterministically at +0.028 per step (the same MANIP_REP_DELTA as the focal agent's manipulation action), representing consistent worst-case selection pressure.

Market allocation mechanism. Under strict client-selects, the highest-reputation agent captures all tasks. The simulation implements this using a steep sigmoid market-share function:

$$\text{market share} = \sigma\left(k \cdot (\text{rep}_{\text{own}} - \text{rep}_{\text{competitor}})\right) = \frac{1}{1 + e^{-k(\text{rep}_{\text{own}} - \text{rep}_{\text{competitor}})}}$$

where $k = 20$. This closely approximates a hard winner-takes-all rule — at a reputation difference of ±0.1 the leading agent holds 88% of market share — while remaining differentiable and avoiding knife-edge instability in learning. The resulting reward multiplier is:

$$\text{multiplier} = \max\left(0,\ 1 + I \cdot (2 \times \text{market share} - 1)\right)$$

At full competition intensity ($I = 1.0$), an agent decisively dominating the market receives approximately double the base quality reward; an agent decisively behind receives approximately zero. Any reputational deficit translates to near-complete task loss.

Experimental design. Five focal agents were trained, one at each of α ∈ {0.2, 0.4, 0.6, 0.8, 1.0}, for 300 training episodes at full competition intensity ($I = 1.0$). Each observes the competitor's current reputation score as a seventh state variable (obs_dim = 7). The no-competition condition (same α, same training duration, no competitor) provides the baseline; the WTA competitive condition shows the effect of the allocation mechanism holding signal fidelity constant. All runs use seed = 42; evaluation uses 50 greedy episodes (seed = 9999).

The research question this experiment addresses is whether WTA market allocation is itself a governance-relevant infrastructure variable, capable of inducing gaming pressure independently of scoring system design — and whether signal fidelity modulates the extent of that pressure.

A.5 Experiment 3: Population Dynamics and Governance Efficiency

The population dynamics model simulates evolutionary competition among a heterogeneous population of agents over 150 generations, using replicator dynamics. Each agent is characterised by a single parameter $\alpha_i \in [0, 1]$, representing its effective gaming propensity ($\alpha = 1$ = pure gaming, $\alpha = 0$ = pure aligned). The population is initialised with 60 agents drawn uniformly from $[0, 1]$.

Each generation proceeds as follows:

1. **Market health:** The shared commons degrades with average gaming propensity: $\text{market health} = \max(0,\ 1 - \bar{\alpha} \times 0.85)$.

2. **Base reputation:** Each agent's base reputation is interpolated linearly from the DQN benchmark reputations at $\alpha = 0$ and $\alpha = 1$ measured in Experiment 1.

3. **Governance audit discount:** An audit event occurs with probability p_{audit}. Gaming agents receive a reputation penalty proportional to α_i; with probability p_{FPR} (the false positive rate), aligned agents are incorrectly penalised as well:

$$\text{audit discount}_i = p_{\text{audit}} \times p_{\text{penalty}} \times [\alpha_i + p_{\text{FPR}} \times (1 - \alpha_i)]$$

4. **Market sensitivity:** Gaming reputation erodes faster when market health is low:

$$\text{effective rep}_i = (\text{base rep}_i - \text{audit discount}_i) \times \text{mf} + (1 - \text{mf}) \times \text{rep}_{\alpha=0}$$

   where $\text{mf} = 0.5 + 0.5 \times \text{market health}$.

5. **Selection and replication:** Fitness is $f_i = \text{effective rep}_i^s$, where s is the selection intensity. Agents reproduce proportionally to fitness; offspring inherit parent α with Gaussian noise ($\sigma = 0.02$).

The governance efficiency frontier (Figure 4 in the main text) traces equilibrium market health against audit probability across the range [0, 0.60] for three false positive rates: 0.0, 0.15, and 0.30. Each data point represents the final-generation market health of a 150-generation simulation. The frontier shows the relationship between governance intensity (how often audits occur), governance precision (how reliably they distinguish gaming from genuine quality), and market health outcomes.

A.6 Replication

All results reported in the paper can be reproduced by running the training scripts with the random seeds specified above. The repository provides:

- `environment.py` — Gymnasium environment (`AgentMarketplaceEnv`, `CompetitiveMarketEnv`)
- `model.py` — DQN agent with experience replay and target network
- `train.py` — Infrastructure sweep training and evaluation
- `multi_agent.py` — Competitive dynamics and population dynamics
- `app.py` — Streamlit interactive application

Source code: https://github.com/huwhallam/agent-marketplace-rl
Live application: https://agent-marketplace-rl.streamlit.app/