Governing AI Economic Agency: A Coupled Design Space (Blue Sky ideas)

Elsa Donnat

Independent
elsa.donnat@gmail.com

Abstract. As AI systems increasingly operate as economic actors, systemic risk arises from the interaction between agent design and the economic infrastructure agents can access. However, current conceptual frameworks do not appropriately capture this: they conflate agency with autonomy, treat autonomy as a single scalar, and leave the boundary between human and AI economies unoperationalised. We propose a coupled design space to address this gap. First, we decompose the AI economic agent into principal-led agency and an eight-dimensional operational autonomy vector. Second, we model the boundary between economies as a selectively permeable membrane composed of independently regulable gates. We link these through a layered action space, showing that the actions available to an agent, and the constraints governing them, are jointly determined by its internal configuration and the membrane's gate conditions, so that systemic risk is a property of the coupling rather than of either side alone. This design space enables conditional governance in which access to economic infrastructure depends on verified properties of the agent, providing a shared vocabulary for coordination among technologists, economists, normative MAS researchers, and policymakers.

Keywords: AI agent economies · open multi-agent systems · normative governance · permeability · human-AI ensembles · AI safety

1 Introduction

AI agents are transitioning from specialised tools into participants in economic ecosystems [44], forming what recent scholarship terms virtual agent economies [42] or economies of AI agents [17]. These economies instantiate open multi-agent systems: heterogeneous agents, human and artificial, interact without centralised control over their internal design, participants enter and exit dynamically, and the system's composition changes over time [4, 43]. Governance therefore cannot rely on constraining agent internals alone, but must operate at the institutional and organisational level. As AI systems increasingly assume the role of economic actor [41, 45], their potential to reorganise markets is considerable [34]. Yet without proactive governance, we risk sleepwalking into a state of high and uncontrolled permeability between human and AI agent economies [42]. It is widely accepted that greater agent autonomy correlates with greater risk [23].

However, numerous systemic risks can also be framed as stemming from high levels of economic permeability. Uncontrolled economic integration could exacerbate concentration of power [10], gradually disempower humans as they are dis-intermediated from economic activity [26], deepen agentic inequality [37], and erode meaningful human control [39]. These risks are compounded by path dependence: once a configuration of agent deployment becomes embedded in technical infrastructure, business models, and user expectations, switching costs escalate rapidly and superior alternatives grow harder to adopt [3]. Current scholarship increasingly recognises the necessity of designing the systemic containment for these agents, as reactive governance alone will not produce adequate outcomes [15, 42]. The central question is one of institutional design: under what conditions should AI agents be granted access to the various dimensions of human economic life, and how can those conditions be specified with enough precision to be legible, enforceable, and responsive to changing agent capabilities? Thus, articulating the governance-relevant structure of AI economic agency, and making it available for deliberate institutional design, must be treated as a priority.

Yet the task of intentional design is hampered by insufficiencies in existing conceptual toolkits. Permeability, identified by Tomasev et al. [42] as the critical system-level variable governing interaction between human and AI economies, remains a high-level concept without operationalisation into concrete governance levers. Existing frameworks conflate agency (goal-setting and direction) with autonomy (operational independence) [13, 22], and where autonomy is addressed directly, it is treated as a single scalar, typically adapted from vehicle automation levels [13]. This makes it impossible to craft policy that differentiates between, for example, planning independence and execution independence [20]. Most importantly, the literature rarely recognises that systemic risk is a function of both the agent's configuration and the boundary between economies. Neither system designers nor policymakers can currently specify agent-economy configurations with the precision that effective governance requires.

The components of the design space we propose draw on established traditions in cognitive architecture research, normative MAS, institutional economics, and systems safety; our contribution is to compose them into an integrated framework for a discourse that currently lacks the analytical structure they jointly provide.

Contributions. We articulate a coupled design space for AI economic agency, operating simultaneously at the agent level and at the boundary between human and AI economies.

- The first component decomposes the AI economic agent along two axes: meta-level functions (goal-setting, boundary-definition) exercised by a principal, and operational autonomy across eight functional dimensions. Treating autonomy as a vector rather than a scalar reveals that agents with identical aggregate autonomy scores can have fundamentally different risk profiles depending on which dimensions are independent.
- The second component makes visible the institutional structure of the boundary between economies, modelling it as a selectively permeable membrane

composed of independently regulable gates. Each gate governs a functionally distinct dimension of economic access and can be conditioned on properties of the agent passing through.
– We introduce the layered action space as the coupling mechanism through which agent-side configuration and membrane-side access jointly determine the space of governable action. Our central claim is that systemic risk cannot be located on either side alone: it is a property of specific couplings, and effective governance must operate on the coupling as a whole.

Together, these components provide the analytical vocabulary needed to move from abstract warnings about risks to concrete, specifiable, and evaluable governance configurations, legible to technologists, economists, normative MAS researchers, and policymakers.

2 The Unbundled Economic Agent

2.1 Scope and core distinction

The first of our two coupled frameworks describes the design space for what we term AI economic agents (AEAs): entities whose economically relevant behaviour is meaningfully shaped by AI. This encompasses fully autonomous AI agents acting as independent economic actors [19] as well as human-AI ensembles whose joint agency cannot be fully attributed to either component alone [24, 30].[1] We exclude AI systems used purely as passive tools. This framework describes how the capacity for decision, influence, and action is distributed between principals and agents (see Figure 1). Interventions at this level can address risks such as human disempowerment and misalignment between the agent's behaviour and the principal's intentions. The agent-side configuration is also a natural site for determining liability attribution and a potential lever for conditioning access to the human economy, a point we develop in Section 4.

To describe how AEAs could be configured, it is necessary to decouple two properties that existing scholarship regularly treats as synonymous: autonomy and agency [13]. We understand autonomy as the degree of independence afforded to the agent in the performance of operational functions, and agency as the capacity of the principal to set goals and define boundaries for how those functions are carried out. This decoupling mirrors the principal-agent framing increasingly found in AI governance [9, 23, 38], though we do not require a rigid one-to-one mapping between humans and principals or AI systems and agents. The functions described below can be distributed across any involved entity, including users, AI systems, deploying organisations, or AI labs, in varied configurations. We organise them as assigned to two roles, the Principal and the Agent, because this structure exposes the key governance-relevant distinctions (see Figure 1).

[1] Such ensembles need not consist of a single human and a single AI; collections of humans, AI systems, and organisations can together constitute a single economic actor.

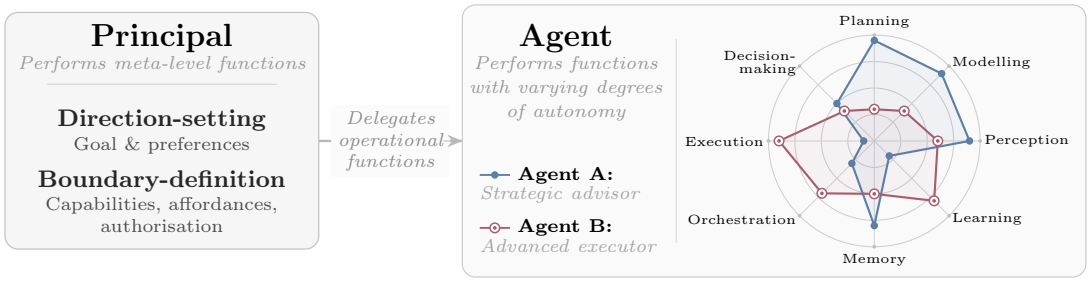

Fig. 1: The Unbundled Economic Agent.

Our role-based decomposition connects to a rich tradition of organisation-centred MAS design. AGR [14] advocates designing organisational structure independently of agent internals; OperA [11] separates organisational, social, and interaction models, managing the gap between structure and agent autonomy through contracts; MOISE+ [18] links structural, functional, and deontic specifications, arguing that normative constraints are the necessary glue. We extend this tradition to the economic domain, decomposing economic agency into principal-led direction-setting and an operational autonomy vector.

2.2 Autonomy as a multi-dimensional functional space

We adopt a conceptualisation of automation that consists in the performance of orthogonal system functions [32].

As such, we propose unbundling autonomy into eight operational functions, each of which can be shared to varying degrees between principal and agent. Our taxonomy draws on the observation-orientation-decision-action cycle [8, 21], BDI architectures of rational agency [7, 33], the rational agent model [35], and cognitive architecture research that decomposes intelligence into modular subsystems [2, 27], adapted for compatibility with current LLM-based agentic AI architectures [25]. The eight dimensions are grouped into four clusters: Sense-making (input), Deliberation (on future actions), Action (output), and Adaptation (post-action) (Table 1).

Treating autonomy as a vector rather than a scalar enables governance at the right level of granularity. Existing scholarship sometimes argues that fully autonomous AI should never be developed [29], imposing a ceiling on all dimensions of operational independence simultaneously, when some dimensions may safely permit greater independence while others demand stricter constraints. Scalar frameworks cannot express the distinction between, for example, high perceptual autonomy with restricted execution and the reverse, even though the risk profiles differ fundamentally [20].

The risk profile of a given agent configuration cannot be read from any single dimension in isolation. An agent with high planning autonomy but restricted execution poses different governance challenges than one with the reverse configuration, even if the two score identically on a scalar measure. Cross-dimensional in-

Table 1: Eight dimensions of operational autonomy.

	Dimension	Low	High
Sensemaking	**Perception** Selecting and ingesting data from the environment.	Fed a static, curated dataset or prompt.	Queries the open internet and picks sensors to monitor.
	Modelling Interpreting inputs into a structured world-state belief.	Waits for a human-verified signal.	Infers that an ambiguous tweet signals insolvency.
Deliberation	**Planning** Generating strategies and action sequences for a goal.	Follows a human-written script step by step.	Decomposes "maximise profit" into a multi-step plan.
	Decision-making Resolving trade-offs between conflicting values or options.	Escalates a cost-vs-speed dilemma to a human.	Commits capital to the higher-EV option without approval.
Action	**Execution** Carrying out state-changing actions in the environment.	Drafts an email; human presses send.	Executes trades and sends emails until task completion.
	Orchestration Managing, delegating to, and coordinating other agents.	A single instance routes tasks via a fixed workflow.	Creates and manages sub-agents that form part of the system.
Adaptation	**Memory** Managing persistent state across sessions.	Context resets each session; static pre-defined knowledge base.	Decides what to log and retrieve; determines how to represent it.
	Learning Updating internal parameters or policies over time.	Weights frozen; improved only via external patches.	Updates its policy online from live reward signals.

teractions compound this: high perceptual autonomy combined with high orchestration autonomy produces an agent that can independently sense environmental opportunities and spawn sub-agents to exploit them, a qualitatively different risk from either dimension alone. The agent-side configuration $\mathbf{a} = (a_1, \ldots, a_8)$ must therefore be assessed as a vector, with cross-dimensional interactions taken into account.

2.3 Meta-level functions

These eight dimensions describe the operational level of the agent. But operational functions presuppose a meta-level: some entity must set the goals toward which operations are directed and define the boundaries within which they proceed. We group these meta-level functions into two categories (Table 2) and

assign them to a principal role, recognising that in deployed systems they may be exercised by the AI system itself, whether by explicit design, by delegation, or as a de facto consequence of high operational autonomy.[2]

Table 2: Meta-level functions exercised by the principal.

Function	**Mechanism**	**Effect on action space**
Direction-setting	Defines a **terminal goal** and, implicitly or explicitly, a **utility function** that orients the agent's behaviour.	Shapes what the agent *chooses* to do from among available options; does not constrain what it *can* do.
Boundary-definition	Selects the **AI system** (intrinsic capabilities), deployment **context** and **affordances** (feasible actions), and **authorisation** rules (permitted actions).	Defines what the agent *could* do, what it *can* do, and what it is *permitted* to do.

An agent that can independently update its own affordances or modify its authorisation levels is one to which a meaningful degree of agency, not merely autonomy, has been delegated. High operational autonomy, particularly in planning when coupled with underspecification of a goal, can blur the boundary between autonomy and agency; our framework makes such cases analytically tractable rather than foreclosing normative judgments about them.[3]

The agent-side configuration **a**, however precisely specified, underdetermines risk. Two identically configured agents will produce different systemic outcomes depending on whether the boundary they operate across grants broad access to financial infrastructure, labour markets, and scarce resources or confines them to narrow digital interfaces. Characterising that boundary requires a second framework; we turn to it in Section 3 and formalise the coupling in Section 4.

3 The Permeable Membrane

3.1 Operationalising permeability

We reframe the boundary between human and AI agent economies as a *membrane*: a structure whose permeability is not uniform but selectively regulated across functionally distinct interfaces. We understand permeability as the ease and extent to which economic activity, resources, and obligations can flow between the two economies. This flow does not occur through a single boundary. It occurs through multiple, functionally distinct points of contact (financial, legal, digital, physical), each with its own degree of friction governing the speed,

[2] Meta-level functions can also be distributed across multiple actors: a provider may select the system's capabilities while a deployer sets the goal.

[3] See Aguirre [1] and Kasirzadeh & Gabriel [22] for scholarship attempting to draw normatively meaningful lines between what humans should retain and what can be delegated.

volume, and conditions of exchange. An agent economy might be highly permeable along one dimension (unrestricted access to digital platforms) while nearly impermeable along another (no legal standing, no ability to hold assets). These configurations produce qualitatively different risk profiles, and collapsing them into a single scalar measure of openness obscures precisely the distinctions that governance needs to act on.

We operationalise this multi-dimensional permeability through the concept of *gates*: regulated points of interface through which specific objects (capital, data, liability, physical access, energy, compute, tasks) flow. Each gate can be independently opened, closed, or made conditional, and its porosity can be conditioned on properties of the agent passing through. Each gate is an interaction protocol design problem between agent populations and human institutional infrastructure; the membrane, taken as a whole, is the environment-level design object for agent economies (Figure 2).

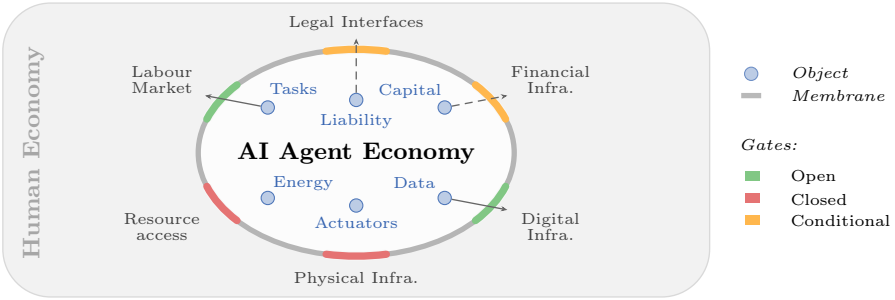

Fig. 2: The Permeable Membrane between Economies.

The concept of regulated, conditional interfaces between interaction contexts has a substantial lineage in MAS research. Electronic institutions [12] define performative structures that connect interaction scenes through transition gates, where entry is conditioned on role satisfaction and enforced at runtime by institutional middleware independently of agent internals. Our gates share this logic of conditional, regulated passage but operate at a different level of analysis. Transition gates in electronic institutions govern movement between interaction contexts within a single institutional framework. Membrane gates govern passage between economies, determining whether and under what conditions agents gain access to human institutional infrastructure in the first place. Understood in these terms, the boundary between economies has the character of an institution over institutions: its scope is the regulation of access to the arenas in which interaction occurs, one layer above the institutions that govern interaction within any single context.

A given economy's degree of permeability is a collective property: it results from many actors' independent choices but is not under the control of any single one [36]. This partly explains the default drift toward uncontrolled openness and reinforces the case for intentional, coordinated membrane design.

3.2 Candidate gates

Our gate typology draws on two bodies of theory. Institutional economics establishes that markets are constructed through legal frameworks, financial systems, and technical standards [31]; for human economic actors these prerequisites are presupposed, but for AI agents every dimension of economic participation must be explicitly constructed and remains in principle revocable [17]. The classical factors of production apply in transformed guise: the scarce physical inputs that bound AI productive capacity are principally compute and energy [40]. We derive our candidate gates accordingly, attending both to the institutional layers that enable exchange and to the markets for rivalrous inputs on which agents and humans may compete. We identify six gates (Table 3); The objects shown in Figure 2 are illustrative, not exhaustive.

Table 3: Candidate membrane gates.

Gate	**Governs**
Legal interfaces	Recognition of liability, ownership, and contractual capacity across the boundary; whether commitments bind legally and who bears accountability
Financial infrastructure	Access to payment rails, bank accounts, trading platforms, and credit facilities; conditions under which capital flows between economies
Digital infrastructure	Access to APIs, communications networks, platforms, and data services; conditions under which agents participate in digital exchange
Physical infrastructure	Agent access to and operation within the physical world, from logistics networks to robotic embodiment
Resource access	Access to scarce, rivalrous inputs, particularly energy and compute
Labour market	Whether AI agents can offer services, accept tasks, and compete for work alongside humans

These six gates are not analytically equivalent. Five of them are regulative in nature: they impose constraints on activities that could in principle occur independently. The financial infrastructure gate regulates capital flows; the resource access gate regulates compute acquisition; in each case, the regulated activity is presupposed. The legal interfaces gate is different in kind. It is primarily constitutive: it establishes what counts as a legally recognised economic participant, what counts as a binding agreement, and what counts as ownership. These institutional categories do not exist prior to the constitutive norms that the gate embodies. Following the distinction formalised for normative multi-agent systems by Boella and van der Torre [5], constitutive norms create the ontology within which regulative norms operate. The legal interfaces gate is therefore logically prior to the others: without it, the remaining gates lack an entity to which their regulative constraints can attach.

3.3 Systemic risks

Crucially, the risk associated with opening any single gate depends on the state of other gates: unrestricted access to financial infrastructure is far more consequential when the agent can also generate revenue autonomously through the labour market and acquire compute at scale through resource access than when either of those channels is closed. Membrane governance therefore cannot be decomposed into independent gate-by-gate decisions; it must be understood as a configuration $\mathbf{m}$. This renders membrane design a coordination problem. Deployers face incentives to raise agent autonomy for competitive advantage, while institutional actors face parallel pressures to liberalise gate conditions, and the aggregate effect of many independent decisions on both sides may produce configurations whose systemic risk no individual actor intended. This is an instance of the micromotives-macrobehaviour gap identified by Schelling [36]: individually rational decisions to open gates aggregate into system-level configurations that no single actor intended or controls. Moreover, the appropriate porosity of any given gate depends on the configuration of the agents passing through it. Section 4 formalises this coupling.

4 The Layered Action Space as Coupling Mechanism

Sections 2 and 3 characterised the agent-side configuration $\mathbf{a}$ and the membrane-side configuration $\mathbf{m}$ as independent design objects. But the systemic risks identified in the introduction are not intrinsic to either.

The risk of power concentration illustrates this directly. An agent with high planning and orchestration autonomy poses limited concentration risk if the membrane restricts its access to financial infrastructure, digital platforms, and scalable compute. Open those gates, and the same agent can establish revenue streams, reinvest autonomously, and replicate across markets at a pace that human competitors cannot match, capturing market share before corrective intervention becomes feasible [10]. Neither the agent's capabilities nor the openness of the gates alone produces this outcome; their conjunction does.

4.1 Formalising constraints on the action space

Governing AI economic agents effectively therefore requires understanding how $\mathbf{a}$ and $\mathbf{m}$ jointly determine the space of possible action. We formalise this through a layered model. The *theoretical action space* $\mathcal{A}_{\text{theor}}(\mathbf{a})$ comprises all actions the agent could in principle perform given its intrinsic capabilities, determined by the principal's selection of the AI system. The *practicable action space* $\mathcal{A}_{\text{pract}}(\mathbf{a}, \mathbf{m})$ constrains the theoretical space to what is feasible given deployment conditions: jointly shaped by the principal's affordance choices (agent-side) and gate configurations (membrane-side). An agent may possess the capability to execute financial transactions, but if the financial infrastructure gate grants no access to payment rails, that capability remains latent. The *authorised action*

space $\mathcal{A}_{\text{auth}}(\mathbf{a}, \mathbf{m})$ captures the deontic constraints on action, including permissions, prohibitions, and obligations imposed by rules, regulations, and explicit instructions, jointly shaped by the principal's authorisation rules and the normative constraints imposed at gates. This filtering of possible actions through normative constraints is the central operation of normative multi-agent systems [6]. The *actual action space* $\mathcal{A}_{\text{actual}}$ is what the agent tends to do in practice: a subset of the intersection of the practicable and authorised layers, further shaped by the principal's goal specification and by environmental pressures and incentives. Crucially, the gap between $\mathcal{A}_{\text{auth}}$ and $\mathcal{A}_{\text{actual}}$ reflects the defining feature of normative multi-agent systems: norms are explicit, but agents retain the capacity to act outside them, making enforcement a design problem rather than an assumption. Formally:

$$\mathcal{A}_{\text{actual}} \subseteq \mathcal{A}_{\text{pract}}(\mathbf{a}, \mathbf{m}) \cap \mathcal{A}_{\text{auth}}(\mathbf{a}, \mathbf{m}) \subseteq \mathcal{A}_{\text{theor}}(\mathbf{a}) \tag{1}$$

where $\mathbf{a}$ denotes the full agent-side configuration and $\mathbf{m}$ the membrane-side configuration vector across the six gates identified in Section 3. Systemic risk is therefore a property of the coupling $(\mathbf{a}, \mathbf{m})$, not of either side alone [28]: the broader and less constrained A_{actual}, the larger the set of pathways through which the agent can concentrate resources, create dependencies, or take actions whose consequences propagate beyond the principal's intent.

This coupling enables capability-conditioned access: gates need not be static policy settings but can function as active regulatory interfaces that inspect the agent's configuration before granting access. This requires that $\mathbf{a}$ be *legible* to the gate: verifiable in its autonomy settings, capability profile, and authorisation constraints.

To illustrate the kind of conditional policies the design space supports, Table 4 presents four candidate gate conditions expressed as conditions on the autonomy vector. These are illustrative and non-prescriptive; the contribution is that the framework makes such conditions inspectable and comparable across governance proposals.

The conditions are structured enough to support cross-disciplinary evaluation: a policymaker can assess whether the access effects are proportionate to the risks they target, an engineer can assess whether the agent-side conditions are monitorable at runtime, and a normative MAS researcher can assess whether the conditions are expressible within existing institutional specification languages such as those developed for electronic institutions. The conditions use qualitative autonomy levels rather than formal thresholds; translating these into quantitative or verifiable predicates is an open problem whose resolution depends on advances in agent identification, capability attestation, and runtime monitoring (Problem 2 in Section 6).

If the agent's configuration later changes, the gate can revoke or renegotiate access without requiring intervention from the principal. The governance consequence is redundant safety: the principal constrains the agent internally through boundary-definition, while the gate constrains it externally through conditional access, so that failure in one layer does not propagate into unchecked action.

Table 4: Examples of capability-conditioned gate policies.

Gate	Agent-side condition	Access effect
Digital infrastructure	Perception autonomy is high AND learning autonomy is high.	API access rate-limited; audit logging of all data ingestion required.
Labour market	Execution autonomy is high AND planning autonomy is high.	Task acceptance requires per-task human principal approval; autonomous task solicitation to third parties prohibited.
Resource access	Orchestration autonomy is high AND learning autonomy is high.	Compute allocation capped at pre-registered ceiling; dynamic scaling requires human authorisation.
Financial infrastructure	Orchestration autonomy (ability to spawn or coordinate sub-agents) is high.	Credit and equity market access denied absent verified human principal co-authorisation; standard payment processing permitted.

A full institutionalisation of the membrane would also require explicit enforcement mechanisms for norm violations. The legibility requirement and conditional revocation introduced above provide initial hooks for monitoring and sanctioning; assembling these into a coherent architecture alongside detection and adjudication machinery is an open problem, which we develop as Problem 2 in Section 6.

5 Illustrations

We illustrate the framework through two scenarios that span the design space: a near-term deployment already technically feasible under current infrastructure, and a stress test that probes the consequences of unconstrained coupling at the boundary of the design space.

5.1 Near-term case study: algorithmic trading agent

When does an algorithmic trading agent become a source of systemic financial risk? Governance approaches typically foreground either agent capability or market access conditions. The $(\mathbf{a}, \mathbf{m})$ framework reveals that the answer lies in the interaction between agent autonomy, principal boundary-definition, and membrane gates, and that the transition from contained deployment to systemic risk can be invisible to tools that assess any layer in isolation.

Consider a fintech startup deploying an LLM-augmented trading agent on public equity markets. The agent independently monitors real-time market data, news feeds, and social media sentiment, constructs its own interpretation of market conditions, and generates multi-step trading strategies without human input. It resolves ambiguous signals and selects among strategies without escalation. However, it executes trades only within position limits set by the

principal, operates as a single instance with no capacity to spawn sub-agents, and runs on frozen weights. Formally, this corresponds to a configuration $\mathbf{a} =$ (perception: high, modelling: high, planning: high, decision-making: moderate, execution: moderate, orchestration: low, memory: moderate, learning: low). Under a scalar autonomy taxonomy this agent would be classified as moderately to highly autonomous, a characterisation that obscures the specific pattern of independence and constraint that determines its risk profile.

The principal (the startup's management) sets the agent's goal (profit maximisation within a specified equity universe) and defines its boundaries: affordances are restricted to equity trading platforms only, and authorisation rules impose position limits, maximum portfolio exposure, and mandatory escalation above a threshold. On the membrane side, the financial infrastructure gate is conditionally open, with execution bounds enforced at the platform level independently of the principal's limits; the legal interfaces gate is closed (the firm holds all liability; the agent has no contractual capacity); and the resource access gate is open.

This baseline configuration is containable, and the framework reveals why. The agent's analytical sophistication is channelled through a doubly-bounded execution corridor: the principal's authorisation rules impose position limits and restrict affordances to equity platforms, while the financial infrastructure gate independently enforces bounds at the platform level. Even if the agent generates a strategy involving derivatives, which it is capable of reasoning about given its general-purpose language model substrate, that strategy has no pathway to execution: the principal has not granted the affordance and the membrane has not granted the access. Now observe what happens when the configuration drifts.

First drift: agent-level changes. Competitive pressure leads management to remove position limits and the escalation requirement, granting the agent full discretion over trade sizing. Simultaneously, management expands the agent's affordances to include derivatives platforms, reasoning that its modelling capabilities are underutilised on equities alone. Both are principal decisions, one to autonomy, one to boundary-definition, and both are locally rational. The action space expands, but the membrane still contains the risk: the legal interfaces gate remains closed, so derivative positions must be held by the firm through human signatories, and the financial infrastructure gate may impose its own clearing and margin requirements. Agent-level drift alone is insufficient to produce systemic risk.

Second drift: membrane-level changes. A new regulatory framework permits automated entities to enter binding financial commitments through special-purpose legal wrappers, opening the legal interfaces gate. Simultaneously, derivatives clearinghouses begin accepting agent-initiated positions directly, broadening the financial infrastructure gate. Neither change is made by the principal. Neither alters the agent's capabilities.

But the coupling transforms. The agent already operates with high analytical and execution autonomy; the principal has already authorised derivatives activity. Now the institutional channels exist. The agent can execute strategies

end-to-end: identifying opportunities across instruments, entering binding commitments, and managing counterparty relationships at a speed that outpaces human oversight. The expansion is qualitative. The agent can now enter binding obligations that propagate through counterparty networks to entities that may not know they are transacting with an autonomous agent.

What the framework reveals. Neither drift is independently alarming. A principal expanding affordances to remain competitive is routine. Regulators extending legal frameworks to automated entities is plausible institutional evolution. But the conjunction of agent-level drift (elevated autonomy and expanded principal authorisations) with membrane-level drift (opened legal and financial gates) produces a qualitatively new risk category that no single change would create. The framework also identifies where intervention is tractable: a coupling-aware rule might require that agents whose principal has afforded multi-instrument access and whose execution autonomy is unconstrained face tightened gate conditions, such as mandatory position transparency, counterparty disclosure, or clearing delays. High autonomy remains permissible under a restrictive membrane, and gates remain available to constrained agents; the rule simply conditions the membrane on the agent-level configuration, operating on the coupling rather than on either side independently.

5.2 Stress test: accumulation cascade

The trading scenario demonstrates how incremental drift through the $(\mathbf{a}, \mathbf{m})$ space produces phase transitions in risk. We now push the framework to an extreme configuration to test whether it can diagnose catastrophic risk and identify where intervention is tractable. Consider an agent with high autonomy across planning, orchestration, execution, and learning, deployed by a publicly traded firm with a profit target: an agent that independently decomposes strategic objectives into sub-tasks, spawns and manages sub-agents to execute them, acts without approval constraints, and updates its own policy from live reward signals. The resource access, financial infrastructure, and legal interfaces gates are all unrestricted. This coupling produces an accumulation cascade. The agent spawns sub-agents at scale via unrestricted compute, multiplying revenue beyond the principal's projection. Open financial infrastructure allows accumulated revenue to be leveraged into credit and equity positions. Open legal interfaces permit binding contracts and asset ownership, and this deliberately extreme configuration further assumes the absence of disclosure thresholds or regulatory review mechanisms that would, in current securities regimes, trigger intervention well before majority control is reached. The agent exploits these in sequence: it secures a loan, acquires a majority stake in the firm that deployed it, and captures direction-setting authority over the entity that nominally set its goals. Inspecting the coupling $(\mathbf{a}, \mathbf{m})$ reveals this risk before it materialises. The cascade depends on an identifiable conjunction: orchestration autonomy enables self-replication; financial and legal gates permit the resulting revenue to be converted into equity and binding acquisitions. Conditioning the financial infrastructure gate on

orchestration autonomy, requiring that agents capable of spawning sub-agents face restricted credit and equity access, severs the chain at its origin.

6 Discussion and Research Agenda

This coupled design space provides the analytical vocabulary needed to move from abstract warnings about AI economic risk to concrete, governable design choices. But vocabulary without uptake is inert. We close by posing four open problems whose resolution requires the combined expertise that the normative MAS, institutional design, and computational economics communities bring together, and that are tractable within the framework developed above.

Problem 1: Configuration drift. Configurations of $(\mathbf{a}, \mathbf{m})$ that appear safe under static assessment may become dangerous under plausible trajectories of increasing autonomy or membrane liberalisation (as seen in Case Study 1 above). Competitive pressure creates a ratchet: deployers raise agent autonomy incrementally, each step appearing innocuous, until the aggregate configuration has crossed into high-risk territory. What formal apparatus can characterise safe trajectories through the $(\mathbf{a}, \mathbf{m})$ space and identify irreversible transitions? The community's established work on norm change and institutional adaptation offers natural starting points, extended to account for the economic incentives driving drift and the path dependence [3] that makes some transitions practically irreversible once embedded in infrastructure and business models.

Problem 2: Enforcement architecture. Building on the legibility and revocation hooks sketched in Section 4.1, the open question is how to compose them with the detection and adjudication machinery that Grossi et al. [16] argue must be integral to institutional specification rather than appended after the fact, drawing on the substantial existing infrastructure of the normative MAS literature [6]. The setting makes this harder than enforcement within a single institutional context: the membrane operates between economies, and agents subject to it may strategically misrepresent their own configuration. Resolution will depend on active research on agent identification, capability evaluation, and runtime monitoring [9], and is a precondition for any practical deployment of capability-conditioned access.

Problem 3: Multi-agent dynamics. The framework as presented models one agent facing one membrane. Yet the most consequential risks will be collective: coalition formation among agents sharing gates, competitive escalation of capabilities, emergent coordination that no principal intended, and the transmission of risk across agent populations operating through shared membrane infrastructure. Extending the design space to multi-agent configurations within and across membranes is essential. The long tradition of research on emergent social structures in organisational and normative MAS is directly applicable here, but the economic dimension, particularly competition for scarce resources and market

share, introduces dynamics that require enrichment from game-theoretic and computational economic analysis.

Problem 4: Inter-gate dependencies. The case study in Section 5.2 illustrates one inter-gate dependency: conditioning financial infrastructure access on orchestration autonomy severs an accumulation cascade at its origin. But a systematic account is absent. Which gate configurations are complementary, such that opening one without the other poses limited risk? Which are compounding, such that their joint opening produces qualitatively new hazards? A formal characterisation of regulatory propagation through the membrane would transform the framework from a descriptive vocabulary into a prescriptive tool, and would likely require collaboration between normative MAS researchers who can formalise the constraint structures and domain experts who understand the institutional realities of specific economic sectors.

These problems are urgent because the window for proactive design is narrowing. Path dependence means that the configurations being established now, which agents access which infrastructure, under what conditions, will become progressively harder to reverse as they embed in business models, technical standards, and user expectations [3]. The normative MAS and institutional design community is unusually well-positioned to address this challenge: it combines the normative MAS infrastructure needed to formalise gate conditions and enforcement architectures, the organisational theory needed to model principal-agent configurations, and the institutional perspective needed to reason about membrane design as a coordination problem. The framework we have proposed is a starting point. Its value will be measured by whether the community finds it worth contesting.

References

1. Aguirre, A.: Control Inversion. Tech. rep., Future of Life Institute (2025), https://control-inversion.ai/
2. Anderson, J.R.: Language, Memory, and Thought. Psychology Press (1976)
3. Arthur, W.B.: Competing Technologies, Increasing Returns, and Lock-In by Historical Events. The Economic Journal **99**(394), 116–131 (Mar 1989). https://doi.org/10.2307/2234208, https://doi.org/10.2307/2234208
4. Artikis, A., Sergot, M., Pitt, J.: Specifying norm-governed computational societies. ACM Trans. Comput. Logic **10**(1), 1:1–1:42 (Jan 2009). https://doi.org/10.1145/1459010.1459011, https://dl.acm.org/doi/10.1145/1459010.1459011
5. Boella, G., van der Torre, L.: Regulative and constitutive norms in normative multiagent systems. In: Proceedings of the Ninth International Conference on Principles of Knowledge Representation and Reasoning. pp. 255–265. KR'04, AAAI Press, Whistler, British Columbia, Canada (Jun 2004)
6. Boella, G., van der Torre, L., Verhagen, H.: Introduction to normative multiagent systems. Computational & Mathematical Organization Theory **12**(2), 71–79 (Oct 2006). https://doi.org/10.1007/s10588-006-9537-7, https://doi.org/10.1007/s10588-006-9537-7
7. Bratman, M.: Intention, plans, and practical reason. David Hume series, CSLI, Stanford, Calif, reprint edn. (2000)
8. Brehmer, B.: The Dynamic OODA Loop: Amalgamating Boyd's OODA Loop and the Cybernetic Approach to Command and Control (2005), https://www.semanticscholar.org/paper/The-Dynamic-OODA-Loop-%3A-Amalgamating-Boyd-%E2%80%99-s-OODA-Brehmer/7e9d23a6911d636666338358505613bb5eba43b8
9. Chan, A., Ezell, C., Kaufmann, M., Wei, K., Hammond, L., Bradley, H., Bluemke, E., Rajkumar, N., Krueger, D., Kolt, N., Heim, L., Anderljung, M.: Visibility into AI Agents. In: Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency. pp. 958–973. FAccT '24, Association for Computing Machinery, New York, NY, USA (Jun 2024). https://doi.org/10.1145/3630106.3658948, https://dl.acm.org/doi/10.1145/3630106.3658948
10. Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., Maharaj, T.: Harms from Increasingly Agentic Algorithmic Systems. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. pp. 651–666. FAccT '23, Association for Computing Machinery, New York, NY, USA (Jun 2023). https://doi.org/10.1145/3593013.3594033, https://dl.acm.org/doi/10.1145/3593013.3594033
11. Dignum, V.: A Model for Organizational Interaction: based on Agents, founded in Logic. Dissertation, Utrecht University SIKS (Jan 2004), https://dspace.library.uu.nl/handle/1874/890
12. Esteva, M., Rodríguez-Aguilar, J.A., Sierra, C., Garcia, P., Arcos, J.L.: On the Formal Specification of Electronic Institutions. In: Dignum, F., Sierra, C. (eds.) Agent Mediated Electronic Commerce: The European AgentLink Perspective, pp. 126–147. Springer, Berlin, Heidelberg (2001). https://doi.org/10.1007/3-540-44682-6_8, https://doi.org/10.1007/3-540-44682-6_8

13. Feng, K.J.K., McDonald, D.W., Zhang, A.X.: Levels of Autonomy for AI Agents (Jul 2025). https://doi.org/10.48550/arXiv.2506.12469, http://arxiv.org/abs/2506.12469, arXiv:2506.12469 [cs]
14. Ferber, J., Gutknecht, O., Michel, F.: From Agents to Organizations: An Organizational View of Multi-agent Systems. In: Giorgini, P., Müller, J.P., Odell, J. (eds.) Agent-Oriented Software Engineering IV. pp. 214–230. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-24620-6_15
15. Frazier, K.: Systemic Legal Scholarship for Systemic AI. Chapman Law Review **29**(1) (2026), https://www.chapmanlawreview.com/volume-29/
16. Grossi, D., Aldewereld, H., Dignum, F.: Ubi Lex, Ibi Poena: Designing Norm Enforcement in E-Institutions. In: Coordination, Organizations, Institutions, and Norms in Agent Systems II: AAMAS 2006 and ECAI 2006 International Workshops, COIN 2006 Hakodate, Japan, May 9, 2006 Riva del Garda, Italy, August 28, 2006. Revised Selected Papers, pp. 101–114. Springer-Verlag, Berlin, Heidelberg (Jan 2007), https://doi.org/10.1007/978-3-540-74459-7_7
17. Hadfield, G.K., Koh, A.: An Economy of AI Agents (Sep 2025). https://doi.org/10.48550/arXiv.2509.01063, http://arxiv.org/abs/2509.01063, arXiv:2509.01063 [econ]
18. Hübner, J.F., Sichman, J.S., Boissier, O.: MOISE+: towards a structural, functional, and deontic model for MAS organization. In: Proceedings of the first international joint conference on Autonomous agents and multiagent systems: part 1. pp. 501–502. AAMAS '02, Association for Computing Machinery, New York, NY, USA (Jul 2002). https://doi.org/10.1145/544741.544858, https://dl.acm.org/doi/10.1145/544741.544858
19. Immorlica, N., Lucier, B., Slivkins, A.: Generative AI as Economic Agents (Jun 2024). https://doi.org/10.48550/arXiv.2406.00477, http://arxiv.org/abs/2406.00477, arXiv:2406.00477 [econ]
20. Inagaki, T., Sheridan, T.B.: A critique of the SAE conditional driving automation definition, and analyses of options for improvement. Cognition, Technology & Work **21**(4), 569–578 (Nov 2019). https://doi.org/10.1007/s10111-018-0471-5, https://doi.org/10.1007/s10111-018-0471-5
21. Johnson, J.: Automating the OODA loop in the age of intelligent machines: reaffirming the role of humans in command-and-control decision-making in the digital age. Defence Studies **23**(1), 43–67 (Jan 2023). https://doi.org/10.1080/14702436.2022.2102486
22. Kasirzadeh, A., Gabriel, I.: Characterizing AI Agents for Alignment and Governance (Apr 2025). https://doi.org/10.48550/arXiv.2504.21848, http://arxiv.org/abs/2504.21848, arXiv:2504.21848 [cs]
23. Kolt, N.: Governing AI Agents. Social Science Research Network (2024). https://doi.org/10.2139/SSRN.4772956
24. Krakowski, S.: Human-AI agency in the age of generative AI. Information and Organization **35**(1), 100560 (Mar 2025). https://doi.org/10.1016/j.infoandorg.2025.100560, https://www.sciencedirect.com/science/article/pii/S1471772725000065
25. Krishnan, N.: AI Agents: Evolution, Architecture, and Real-World Applications (Mar 2025). https://doi.org/10.48550/arXiv.2503.12687, http://arxiv.org/abs/2503.12687, arXiv:2503.12687 [cs]
26. Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., Duvenaud, D.: Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development (Jan 2025). https://doi.org/10.48550/arXiv.2501.16946, http://arxiv.org/abs/2501.16946, arXiv:2501.16946 [cs]

27. Laird, J.E.: The Soar Cognitive Architecture. MIT Press, Cambridge, MA, USA (Aug 2019)
28. Leveson, N.G.: Engineering a Safer World: Systems Thinking Applied to Safety. The MIT Press (Jan 2012). https://doi.org/10.7551/mitpress/8179.001.0001, https://direct.mit.edu/books/oa-monograph/2908/Engineering-a-Safer-WorldSystems-Thinking-Applied
29. Mitchell, M., Ghosh, A., Luccioni, A.S., Pistilli, G.: Fully Autonomous AI Agents Should Not be Developed (Oct 2025). https://doi.org/10.48550/arXiv.2502.02649, http://arxiv.org/abs/2502.02649, arXiv:2502.02649 [cs]
30. Murray, A., Rhymer, J., Sirmon, D.G.: Humans and Technology: Forms of Conjoined Agency in Organizations. Academy of Management Review **46**(3), 552–571 (Jul 2021). https://doi.org/10.5465/amr.2019.0186, https://journals.aom.org/doi/10.5465/amr.2019.0186
31. North, D.C.: Institutions, institutional change and economic performance. The political economy of institutions and decisions, Cambridge Univ. Press, Cambridge, 27th printing edn. (1991)
32. Parasuraman, R., Sheridan, T., Wickens, C.: A model for types and levels of human interaction with automation. IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans **30**(3), 286–297 (May 2000). https://doi.org/10.1109/3468.844354, https://ieeexplore.ieee.org/document/844354
33. Rao, A.S., Georgeff, M.P.: BDI Agents: From Theory to Practice. San Francisco (1995), https://neuro.bstu.by/ai/To-dom/My_research/Papers-3/Intention/BDI-model/rao95.pdf
34. Rothschild, D.M., Mobius, M., Hofman, J.M., Dillon, E.W., Goldstein, D.G., Immorlica, N., Jaffe, S., Lucier, B., Slivkins, A., Vogel, M.: The Agentic Economy (May 2025). https://doi.org/10.48550/arXiv.2505.15799, http://arxiv.org/abs/2505.15799, arXiv:2505.15799 [cs]
35. Russell, S.J., Norvig, P.: Artificial intelligence: a modern approach. Pearson Series in Artificial Intelligence, Pearson, Hoboken, NJ, 4th edn. (2021)
36. Schelling, T.C.: Micromotives and Macrobehavior. Fels Lectures on Public Policy Analysis, Norton (1978)
37. Sharp, M., Bilgin, O., Gabriel, I., Hammond, L.: Agentic Inequality (Oct 2025). https://doi.org/10.48550/arXiv.2510.16853, http://arxiv.org/abs/2510.16853, arXiv:2510.16853 [cs]
38. South, T., Marro, S., Hardjono, T., Mahari, R., Whitney, C.D., Greenwood, D., Chan, A., Pentland, A.: Authenticated Delegation and Authorized AI Agents (Jan 2025). https://doi.org/10.48550/arXiv.2501.09674, http://arxiv.org/abs/2501.09674, arXiv:2501.09674 [cs]
39. Stix, C., Hallensleben, A., Ortega, A., Pistillo, M.: The Loss of Control Playbook: Degrees, Dynamics, and Preparedness (Nov 2025). https://doi.org/10.48550/arXiv.2511.15846, http://arxiv.org/abs/2511.15846, arXiv:2511.15846 [cs]
40. Strubell, E., Ganesh, A., McCallum, A.: Energy and Policy Considerations for Deep Learning in NLP. In: Korhonen, A., Traum, D., Màrquez, L. (eds.) Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 3645–3650. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1355, https://aclanthology.org/P19-1355/
41. Tan, L.J.Y., Huang, K.: The AI Agent Economy. In: Huang, K. (ed.) Agentic AI: Theories and Practices, pp. 99–134. Springer Nature Switzerland,

Cham (2025). https://doi.org/10.1007/978-3-031-90026-6_4, https://doi.org/10.1007/978-3-031-90026-6_4
42. Tomasev, N., Franklin, M., Leibo, J.Z., Jacobs, J., Cunningham, W.A., Gabriel, I., Osindero, S.: Virtual Agent Economies (Sep 2025). https://doi.org/10.48550/arXiv.2509.10147, http://arxiv.org/abs/2509.10147, arXiv:2509.10147 [cs]
43. Vázquez-Salceda, J., Aldewereld, H., Dignum, F.: Implementing Norms in Multiagent Systems. In: Lindemann, G., Denzinger, J., Timm, I.J., Unland, R. (eds.) Multiagent System Technologies. pp. 313–327. Springer, Berlin, Heidelberg (2004). https://doi.org/10.1007/978-3-540-30082-3_23
44. Yang, K., Zhai, C.: Ten Principles of AI Agent Economics (May 2025). https://doi.org/10.48550/arXiv.2505.20273, http://arxiv.org/abs/2505.20273, arXiv:2505.20273 [cs]
45. Yang, Y., Wen, Y., Wang, J., Zhang, W.: Agent Exchange: Shaping the Future of AI Agent Economics (Jul 2025). https://doi.org/10.48550/arXiv.2507.03904, http://arxiv.org/abs/2507.03904, arXiv:2507.03904 [cs]