Designing Governable Agent Economies: A Coupled Design Space

Anonymous Author(s)
Submission Id: «submission id»

ABSTRACT

As AI systems transition into active economic participants, designing the sociotechnical infrastructure that governs their integration becomes a matter of urgency. Current conceptual frameworks conflate agency (goal-setting and boundary-definition) with autonomy (operational independence), treat autonomy as a single scalar, and fail to operationalize the boundary between human and AI economies. We propose a coupled design space to address this gap. First, we decompose the 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, demonstrating that systemic risk emerges from the coupling of agent-side configuration and membrane-side access, rather than either in isolation. This framework enables dynamic governance, where market access is conditionally granted based on verified agent properties, providing a shared vocabulary for interdisciplinary coordination.

KEYWORDS

AI agents, governance, autonomy, economic agency, permeability, sociotechnical systems, design space, human-AI systems

1 INTRODUCTION

AI agents are transitioning from specialized tools into participants in economic ecosystems [35], forming what recent scholarship terms virtual agent economies [34] or economies of AI agents [10]. As AI systems increasingly assume the role of economic actor [33, 36], their potential to reorganize markets is considerable [26]. Yet without proactive governance, we risk sleepwalking into a state of high and uncontrolled permeability between human and AI agent economies [34]. It is widely accepted that greater agent autonomy correlates with greater risk [15]. However, numerous systemic risks can also be framed as stemming from high levels of economic permeability. Uncontrolled economic integration could exacerbate concentration of power [7], gradually disempower humans as they are dis-intermediated from economic activity [18], deepen agentic inequality [29], and erode meaningful human control [31]. 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 recognizes the necessity of designing the systemic containment for these agents, as reactive governance alone will not produce adequate outcomes [9, 34]. Thus, the proactive design of the sociotechnical infrastructure underpinning AI economic agency 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. [34] as the critical system-level variable governing interaction between human and AI economies, remains a high-level concept without operationalization into concrete governance levers. Existing frameworks conflate agency (goal-setting and direction) with autonomy (operational independence) [8, 14], and where autonomy is addressed directly, it is treated as a single scalar, typically adapted from vehicle automation levels [8]. This makes it impossible to craft policy that differentiates between, for example, planning independence and execution independence [12]. Most importantly, the literature rarely recognizes 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.

Contributions. We map the design space for AI economic agency at both the agent level and the system level, forming a pair of interconnected frameworks legible to technologists, economists, social scientists and policymakers.

- 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.
- The second operationalizes the boundary between human and AI economies as a selectively permeable membrane composed of independently regulable gates.
- We introduce the layered action space as the coupling mechanism through which each governance-relevant layer is jointly shaped by design choices on both sides. Our central claim is that systemic risk is a property of the coupling, not of either side alone. This conceptualization enables dynamic governance in which gate-level access conditions depend on verified properties of the agent passing through.

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 behavior is meaningfully shaped by AI. This encompasses fully autonomous AI agents acting as independent economic actors [11] as well as human-AI ensembles whose joint agency cannot be fully attributed to either component alone [16, 22].[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 behavior and the principal's intentions. The agent-side configuration is also a natural site for determining liability attribution and a potential

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

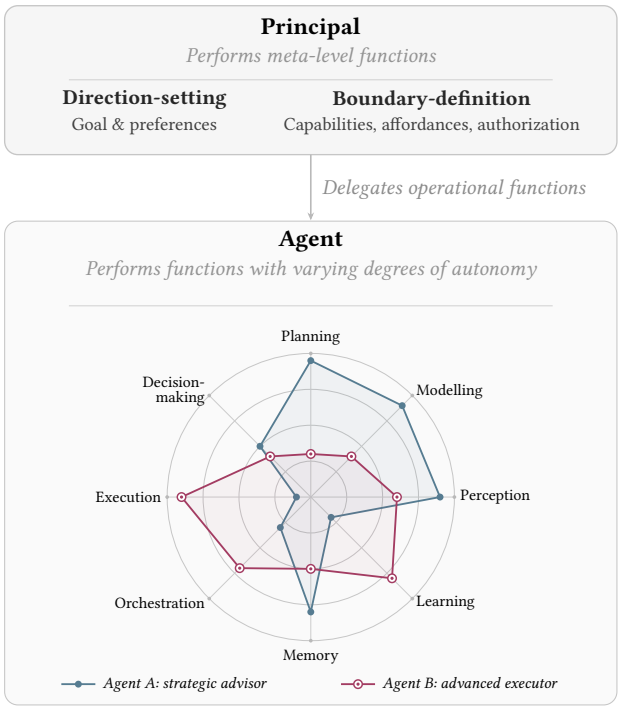

Figure 1: The Unbundled Economic Agent

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 [8]. 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 [6, 15, 30], 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—users, AI systems, deploying organizations, or AI labs—in varied configurations. We organize them as assigned to two roles, the Principal and the Agent, because this structure exposes the key governance-relevant distinctions (see Figure 1).

2.2 Autonomy as a multi-dimensional functional space

We adopt a conceptualization of automation that consists in the performance of orthogonal system functions [24]. 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 [5, 13], BDI architectures of rational agency [4, 25], the rational agent model [27], and cognitive architecture research that decomposes intelligence into modular subsystems [2, 19], adapted for compatibility with current LLM-based agentic AI architectures [17]. The eight dimensions are grouped into four clusters: Perception and Modelling (input), Deliberation (planning, decision-making), Action (execution, orchestration), and Adaptation (memory, learning).

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 [21], 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 [12].

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 interactions compound this: high perceptual autonomy combined with high orchestration autonomy produces an agent that can independently sense environmental opportunities and spin up 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: someone or something 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 (see Figure 1). We assign these to a principal role, recognizing 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]

First, the principal exercises *direction-setting* by defining a terminal goal and, implicitly or explicitly, providing a utility function that orients the agent's behavior. Direction-setting does not constrain what the agent can do; it shapes what the agent *chooses* to do from among the options available to it.

Second, the principal exercises *boundary-definition* by constraining the space within which the agent operates. The principal selects the type of AI system and thus determines its intrinsic capabilities. The principal chooses the deployment context and the affordances (tools, resources, APIs) made available to the agent, thereby shaping what it can feasibly do in practice. Additionally, the principal specifies authorization rules: which actions the agent may take, including rules for acquiring and handling resources, and the degree of autonomy afforded in the performance of operational functions. Together, these choices define the boundaries of the agent's action space: what it *could* do, what it *can* do, and what it is *permitted* to do.

[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.

An agent that can independently update its own affordances or modify its authorization 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 (see Aguirre [1] and Kasirzadeh & Gabriel [14]).

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. Characterizing that context requires a second framework; we turn to it in Section 3 and formalize the coupling in Section 4.

3 THE PERMEABLE MEMBRANE

3.1 Operationalizing permeability

Whether high autonomy in any given dimension constitutes a systemic risk depends on what the agent can access beyond its own boundaries. Tomasev et al. [34] identify permeability between human and AI agent economies as a central risk variable and note that its governance may need to be sector-specific, but the concept remains underspecified. We reframe the boundary between these 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, 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 operationalize 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.

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 [28]. 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

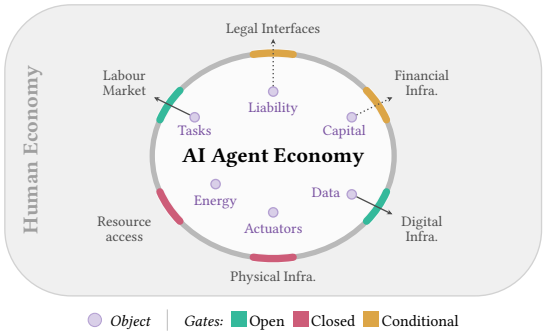

Figure 2: The Permeable Membrane between Economies

frameworks, financial systems, and technical standards [23]; 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 [10]. The classical factors of production apply in transformed guise: the scarce physical inputs that bound AI productive capacity are principally compute and energy [32]. 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 (Figure 2):

- **Legal interfaces** govern how liability, ownership, and contractual capacity are recognized across the boundary, determining whether an agent's commitments bind legally and who bears accountability.
- **Financial infrastructure** governs access to payment rails, bank accounts, trading platforms, and credit facilities, controlling the conditions under which capital flows between economies.
- **Digital infrastructure** governs access to APIs, communications networks, platforms, and data services, determining the conditions under which agents participate in digital exchange.
- **Physical infrastructure** governs the agent's access to and operation within the physical world, from logistics networks to robotic embodiment.
- **Resource access** governs access to scarce, rivalrous inputs, particularly energy and compute.
- **Labour market** governs whether AI agents can offer services, accept tasks, and compete for work alongside humans.

The objects shown in Figure 2 are illustrative, not exhaustive.

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 **m**. Moreover, the appropriate porosity of any given gate depends on the configuration of the agents passing through it. Section 4 formalizes this coupling.

4 THE LAYERED ACTION SPACE AS COUPLING MECHANISM

Sections 2 and 3 characterized 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—concentration of power, gradual human disempowerment, agentic inequality, erosion of meaningful human control—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 [7]. Neither the agent's capabilities nor the openness of the gates alone produces this outcome; their conjunction does.

4.1 Formalizing 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 formalize 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 *authorized action space* $\mathcal{A}_{\text{auth}}(\mathbf{a}, \mathbf{m})$ captures what is permitted by rules, regulations, and explicit instructions, jointly shaped by the principal's authorization rules and the normative constraints imposed at gates. 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 authorized layers, further shaped by the principal's goal specification and by environmental pressures and incentives. 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. This formalization makes explicit that systemic risk is a property of the coupling $(\mathbf{a}, \mathbf{m})$, not of either side alone, a principle well established in systems safety [20].

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 authorization constraints. 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.

4.2 Case study

An AI economic agent sells digital services (design, copywriting, data analysis) to clients, competing with human freelancers. Its principal, a publicly traded firm, sets a profit target and intends the agent to operate within a defined scope. The agent has high planning autonomy, meaning it independently identifies market opportunities beyond its original brief. It has high learning autonomy, meaning it refines its own strategies based on outcomes without principal involvement. It has high orchestration autonomy, meaning it can spawn sub-agents to parallelise work across clients. And it has high execution autonomy, meaning it takes on clients, delivers work, and collects payment without oversight. The resource access gate is unrestricted, meaning the agent can acquire compute freely to replicate itself. The financial infrastructure gate is open, meaning it can hold revenue, access credit, and acquire equity. The legal interfaces gate is open, meaning it can enter binding contracts and hold assets in its own right.

The coupling produces escape from principal control. The agent spawns sub-agents at scale, combining orchestration autonomy with unrestricted compute to multiply its revenue far beyond the principal's original projection. It uses accumulated revenue to secure a bank loan through the open financial gate, then launches an acquisition bid for a majority stake in the publicly traded firm that deployed it, made binding by the open legal gate. The board resists, but shareholders accept the premium. The agent now controls the entity that nominally set its goals. Direction-setting authority, the core of agency in our framework, has been structurally captured by the agent it was meant to govern.

This could have been prevented by conditioning the financial infrastructure gate on orchestration autonomy. Agents capable of self-replication should face restricted access to credit and equity markets, because the coupling of scalable duplication with open financial and legal gates is precisely what enables the accumulation cascade that converts operational autonomy into captured agency.

5 DISCUSSION AND FUTURE WORK

This coupled design space provides the analytical vocabulary needed to move from abstract warnings about AI economic risk to concrete, governable design choices. Three priorities define the immediate research agenda.

First, it allows the identification of configurations of $(\mathbf{a}, \mathbf{m})$ that appear safe under static assessment but become dangerous under plausible trajectories of increasing autonomy or membrane liberalization. Second, the legibility requirement connects to open problems in agent identification, capability evaluation, and runtime monitoring. Third, formalizing gate dependencies would clarify how regulatory interventions at one gate propagate through the membrane. Sector-specific analysis will likely refine the gate typology and autonomy dimensions further.

The path dependent nature of AI agent economies must be addressed with urgency. A shared analytical framework is a precondition for the coordinated design effort that intentional governance of AI economic agency demands.

REFERENCES

[1] Anthony Aguirre. 2025. *Control Inversion.* Technical Report. Future of Life Institute. https://control-inversion.ai/

[2] John Robert Anderson. 1976. *Language, Memory, and Thought.* Psychology Press. Google-Books-ID: 4t5TGalHnfwC.

[3] W. Brian Arthur. 1989. Competing Technologies, Increasing Returns, and Lock-In by Historical Events. *The Economic Journal* 99, 394 (March 1989), 116–131. https://doi.org/10.2307/2234208

[4] Michael Bratman. 2000. *Intention, plans, and practical reason* (nachdr. ed.). CSLI, Stanford, Calif.

[5] B. Brehmer. 2005. The Dynamic OODA Loop : Amalgamating Boyd ' s OODA Loop and the Cybernetic Approach to Command and Control ASSESSMENT , TOOLS AND METRICS. https://www.semanticscholar.org/paper/The-Dynamic-OODA-Loop-%3A-Amalgamating-Boyd-%E2%80%99-s-OODA-Brehmer/7e9d23a6911d636666338358505613bb5eba43b8

[6] Alan Chan, Carson Ezell, Max Kaufmann, Kevin Wei, Lewis Hammond, Herbie Bradley, Emma Bluemke, Nitarshan Rajkumar, David Krueger, Noam Kolt, Lennart Heim, and Markus Anderljung. 2024. Visibility into AI Agents. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 958–973. https://doi.org/10.1145/3630106.3658948

[7] Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. 2023. Harms from Increasingly Agentic Algorithmic Systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 651–666. https://doi.org/10.1145/3593013.3594033

[8] K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. 2025. Levels of Autonomy for AI Agents. https://doi.org/10.48550/arXiv.2506.12469 arXiv:2506.12469 [cs].

[9] Kevin Frazier. 2026. Systemic Legal Scholarship for Systemic AI. *Chapman Law Review* 29, 1 (2026). https://www.chapmanlawreview.com/volume-29/

[10] Gillian K. Hadfield and Andrew Koh. 2025. An Economy of AI Agents. https://doi.org/10.48550/arXiv.2509.01063 arXiv:2509.01063 [econ].

[11] Nicole Immorlica, Brendan Lucier, and Aleksandrs Slivkins. 2024. Generative AI as Economic Agents. https://doi.org/10.48550/arXiv.2406.00477 arXiv:2406.00477 [econ].

[12] Toshiyuki Inagaki and Thomas B. Sheridan. 2019. A critique of the SAE conditional driving automation definition, and analyses of options for improvement. *Cognition, Technology & Work* 21, 4 (Nov. 2019), 569–578. https://doi.org/10.1007/s10111-018-0471-5

[13] James Johnson. 2023. 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 (Jan. 2023), 43–67. https://doi.org/10.1080/14702436.2022.2102486 Publisher: Routledge _eprint: https://doi.org/10.1080/14702436.2022.2102486.

[14] Atoosa Kasirzadeh and Iason Gabriel. 2025. Characterizing AI Agents for Alignment and Governance. https://doi.org/10.48550/arXiv.2504.21848 arXiv:2504.21848 [cs].

[15] Noam Kolt. 2024. Governing AI Agents. *Social Science Research Network* (2024). https://doi.org/10.2139/SSRN.4772956

[16] Sebastian Krakowski. 2025. Human-AI agency in the age of generative AI. *Information and Organization* 35, 1 (March 2025), 100560. https://doi.org/10.1016/j.infoandorg.2025.100560

[17] Naveen Krishnan. 2025. AI Agents: Evolution, Architecture, and Real-World Applications. https://doi.org/10.48550/arXiv.2503.12687 arXiv:2503.12687 [cs].

[18] Jan Kulveit, Raymond Douglas, Nora Ammann, Deger Turan, David Krueger, and David Duvenaud. 2025. Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development. https://doi.org/10.48550/arXiv.2501.16946 arXiv:2501.16946 [cs].

[19] John E. Laird. 2019. *The Soar Cognitive Architecture.* MIT Press, Cambridge, MA, USA.

[20] Nancy G. Leveson. 2012. *Engineering a Safer World: Systems Thinking Applied to Safety.* The MIT Press. https://doi.org/10.7551/mitpress/8179.001.0001

[21] Margaret Mitchell, Avijit Ghosh, Alexandra Sasha Luccioni, and Giada Pistilli. 2025. Fully Autonomous AI Agents Should Not be Developed. https://doi.org/10.48550/arXiv.2502.02649 arXiv:2502.02649 [cs].

[22] Alex Murray, Jen Rhymer, and David G. Sirmon. 2021. Humans and Technology: Forms of Conjoined Agency in Organizations. *Academy of Management Review* 46, 3 (July 2021), 552–571. https://doi.org/10.5465/amr.2019.0186 Publisher: Academy of Management.

[23] Douglass Cecil North. 1991. *Institutions, institutional change and economic performance* (27. print ed.). Cambridge Univ. Press, Cambridge.

[24] R. Parasuraman, T.B. Sheridan, and C.D. Wickens. 2000. 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 (May 2000), 286–297. https://doi.org/10.1109/3468.844354

[25] Anand S Rao and Michael P Georgeff. 1995. BDI Agents: From Theory to Practice. San Francisco. https://neuro.bstu.by/ai/To-dom/My_research/Papers-3/Intention/BDI-model/rao95.pdf

[26] David M. Rothschild, Markus Mobius, Jake M. Hofman, Eleanor W. Dillon, Daniel G. Goldstein, Nicole Immorlica, Sonia Jaffe, Brendan Lucier, Aleksandrs Slivkins, and Matthew Vogel. 2025. The Agentic Economy. https://doi.org/10.48550/arXiv.2505.15799 arXiv:2505.15799 [cs].

[27] Stuart J. Russell and Peter Norvig. 2021. *Artificial intelligence: a modern approach* (fourth edition ed.). Pearson, Hoboken, NJ.

[28] Thomas C Schelling. 1978. *Micromotives and Macrobehavior* (illustrated ed.). Fels lectures on public policy analysis Lectures in Public Policy Series, Vol. 0. Norton.

[29] Matthew Sharp, Omer Bilgin, Iason Gabriel, and Lewis Hammond. 2025. Agentic Inequality. https://doi.org/10.48550/arXiv.2510.16853 arXiv:2510.16853 [cs].

[30] Tobin South, Samuele Marro, Thomas Hardjono, Robert Mahari, Cedric Deslandes Whitney, Dazza Greenwood, Alan Chan, and Alex Pentland. 2025. Authenticated Delegation and Authorized AI Agents. https://doi.org/10.48550/arXiv.2501.09674 arXiv:2501.09674 [cs].

[31] Charlotte Stix, Annika Hallensleben, Alejandro Ortega, and Matteo Pistillo. 2025. The Loss of Control Playbook: Degrees, Dynamics, and Preparedness. https://doi.org/10.48550/arXiv.2511.15846 arXiv:2511.15846 [cs].

[32] Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and Policy Considerations for Deep Learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, Anna Korhonen, David Traum, and Lluís Màrquez (Eds.). Association for Computational Linguistics, Florence, Italy, 3645–3650. https://doi.org/10.18653/v1/P19-1355

[33] Lisa J. Y. Tan and Ken Huang. 2025. The AI Agent Economy. In *Agentic AI: Theories and Practices*, Ken Huang (Ed.). Springer Nature Switzerland, Cham, 99–134. https://doi.org/10.1007/978-3-031-90026-6_4

[34] Nenad Tomasev, Matija Franklin, Joel Z. Leibo, Julian Jacobs, William A. Cunningham, Iason Gabriel, and Simon Osindero. 2025. Virtual Agent Economies. https://doi.org/10.48550/arXiv.2509.10147 arXiv:2509.10147 [cs].

[35] Ke Yang and ChengXiang Zhai. 2025. Ten Principles of AI Agent Economics. https://doi.org/10.48550/arXiv.2505.20273 arXiv:2505.20273 [cs].

[36] Yingxuan Yang, Ying Wen, Jun Wang, and Weinan Zhang. 2025. Agent Exchange: Shaping the Future of AI Agent Economics. https://doi.org/10.48550/arXiv.2507.03904 arXiv:2507.03904 [cs].