NETWORK LAW REVIEW

ARTICLE

Principal-Agent Dynamics and Digital (Platform) Economics in the Age of Agentic AI

Volker Stocker & William Lehr

The Network Law Review is pleased to present a special issue entitled "The Law & Technology & Economics of AI." This issue brings together multiple disciplines around a central question: What kind of governance does AI demand? A workshop with all the contributors took place on May 22–23, 2025, in Hong Kong, hosted by Adrian Kuenzler (HKU Law School), Thibault Schrepel (Vrije Universiteit Amsterdam), and Volker Stocker (Weizenbaum Institute). They also serve as the editors.

***

Abstract: This article applies the principal–agent framework to the use of autonomous AI systems in digital markets. It examines the challenge of aligning AI agents with the interests of end-users, given that many systems may also reflect the objectives of developers, platform providers, or advertisers. These "shadow principals" create persistent information asymmetries and reduce user control. The paper also considers how agent-to-agent transactions could alter the economics of digital platforms by weakening attention-based business models, reshaping network effects, and redistributing bargaining power. The analysis concludes that governance will depend on building measurement systems capable of detecting misalignment and evaluating agentic bargaining outcomes.

1. Introduction

We are getting a glimpse of a trajectory that might take us toward an agentic AI world. With ChatGPT's introduction in November 2022, the potential for digital technology to augment and potentially replace direct human engagement with tasks

flashed into popular consciousness, making many aware for the first time just how pervasive and dependent on digital technologies modern economies and societies already are. Building on the success of Large Language Models (LLMs) and the proliferation of add-on enhancements and applications, businesses across the economy and especially those offering digital applications are rushing to deploy AI agents to augment and enhance the capabilities of the digital technologies they have already been using. From search engines to all kinds of software applications, from customer service to human resource to supply chain management, AI technologies are being utilized to smooth the digital automation of an ever-wider range of tasks.[1]

The Agentic AI future will involve AI tools that are not only more deeply embedded in our social and economic realities but also more proactive.[2] AI agents are expected to act with broad autonomy and limited human supervision, to achieve task-oriented goals predetermined by their principals (which are expected to be end-users, but as we explain further below, the identity of the principal may be unclear). Proposed use cases anticipate AI agents planning and executing increasingly complex, multi-step processes on behalf of their principals, coordinating and negotiating with other AI agents, while learning and adapting from their interactions. As AI agents are widely anticipated to be adaptive, (locally) customized and personalized, and self-evolving (e.g., Gao et al., 2025), it is widely expected that AI systems are increasingly assisting, augmenting, and shaping human users, their knowledge, perceptions and preferences, as well as choices and decisions. All of this is predicated on a fundamental transformation of three critical interaction paradigms: direct human-AI collaboration, interactions between AI-augmented humans, and machine-to-machine communications between AI systems.

In economics, the process of automation may be viewed through the lens of the Principal-Agent Problem (PAP). This framework has proved extremely powerful for evaluating the economic control problem that arises when a Principal (e.g., the state, a business owner, etc.) seeks to accomplish a goal (e.g., maximize social welfare, profits, etc.) by relying on the contracted activity of agents (e.g., firms in an industry under the social contract for regulating the economy, employees of a firm, etc.) who are themselves privately optimizing agents.[3]

---

[1] Numerous industry analysts and spokespersons for leading tech companies dubbed 2025 the year for "Agentic AI" (e.g., Deslandes, 2025; Caserman, 2025; Capgemini, 2025; Shaw, 2025). Note, however, that definitions of AI agents and agentic AI remain somewhat fluid and elusive (e.g., Jarrahi and Ritala, 2025).

[2] See Kapoor et al. (2024). Kasirzadeh and Gabriel (2025) provide a four dimension framework to characterize the diversity of AI agents, focusing on: (i) different levels of agent autonomy (and corresponding human control), (ii) the agent's efficacy and its "ability to interact with and have a causal impact upon the environment" (Kasirzadeh and Gabriel, 2025, p. 8), (iii) varying levels of goal complexity, and (iv) different degrees of generality.

[3] The seminal article in the evolution of the PAP model is Jensen & Meckling (1976). Several good textbooks on the theory and its further development include Milgrom & Roberts (1992), Laffont and

In this essay, we apply the PAP framework to investigate two sets of issues: the challenge of ensuring appropriate alignment between human principals and their AI-agents; and the economic implications of AI-agents differentially augmenting market participants or interacting directly on digital (platform) markets and value chains.

2. Whose Agent? Aligning Human Principal Control and AI Agents

The PAP framework is typically framed as a constrained optimization problem, where the constraints arise because the principal can only imperfectly (monitor and) control the activity of the agents. That is, the contracts are incomplete, with asymmetric information identified as one common cause. The presence of such asymmetric information and the imperfect control it can result in can give rise to a host of system-level problems, such as moral hazard and adverse selection. We refer to these as system-level problems because they arise from the imperfect and costly misalignment of principal and agent goals, which constrains the realization of optimal outcomes that might otherwise be attainable if the constraints were relaxed and the goals better aligned.

At first glance, AI agents appear to fit well into the PAP framework and the alignment problem explained above: users (e.g., individuals or businesses) serve as principals who instruct agents to perform tasks on their behalf. This relationship mirrors traditional economic arrangements where one party delegates authority to another to act in their interest. Although delegation is a common practice in economic organizations and is typically predicated on a hierarchical structure, when delegation involves digital software agents, a wide range of additional complications arises.[4]

For example, in many (most?) cases, individual end-users are unlikely to be the parties that program and deploy the AI agents that are presumed to be acting on their behalf. Whitt (2024a, 2024b) and Schneier (2025) highlight the potential for this to result in misalignment problems where the AI agents act as "double agents," supposedly in service of the user principal while subtly (and potentially disproportionately) prioritizing the interests of one or more other ecosystem stakeholders, including AI agent or model providers, advertisers, or others (which may be referred to as "shadow principals"). These alignment issues are not new and are familiar from other algorithmic systems, such as search engines or social media

Martimort (2002), and Bolton and Dewatripont (2005). For a perspective on recent advances and how PAP relates to what is sometimes referred to as the New Institutional Economics, see Ménard & Shirley (2022).

[4] Multiple experts have explored human-AI agent interactions in a variety of PAP(-related) settings, including Whitt (2024a, 2024b), Schneier (2025), Kolt (2025), Jarrahi and Ritala (2025), and Kasirzadeh and Gabriel (2025).

platforms. What is different with AI agents is the form and depth of human-AI collaboration, as well as the PAP context. While emerging control challenges for the user principals are complex, one key challenge will be that end-users may not be aware of whether and how they are dependent on third-party products, providers, and data (at various stages and in various ways), let alone what such dependency really means for them.[5] This opacity might be exacerbated by limited transparency regarding AI model development and design, as well as the potential for manipulation of the information sources that "feed" the agents (e.g., through the manipulation of model training data or the selective feeding of data to agents via APIs). The latter becomes increasingly relevant when agents autonomously search for and retrieve information with little or no human supervision.[6]

A key concern is that entities may shape the agent's objective function and behavior in ways undetectable, invisible, and incomprehensible to user principals who operate these agents under the illusion of 'exclusive control'. If true, this prevents user principals from responding or taking measures to counter such influence in the first place, e.g., by building or customizing AI models and agents. The opacity risks are exacerbated by the potential for feedback loops and mutual influence between humans and AI systems, especially since agents are assumed to be deeply embedded in the users' lives (Kirk et al., 2025). Absent responses and mitigation efforts, opacity may render persistent information asymmetries more likely, and with this, bias and sycophancy, nudging, or other forms of (targeted and personalized) manipulation.[7] All of these may influence user choices and decisions in ways that are not only undesirable but undetected or undetectable by the users themselves. This may result in dependencies, illusions of trust and control, and reliance, which may cause potential lock-in effects.

While knowing whose interests an AI Agent is actually serving is key to addressing the human principal-AI agent alignment problem, better transparency only addresses part of the challenge. Designing better AI agents confronts multiple challenges. One significant challenge is that much of the focus on improving the performance of AI Agents fails to adequately treat the challenge holistically. For example, Qian et al. (2025) highlight how LLM-based agents align with human strategies and norms in the context of automated negotiation, finding that the focus has been on efficiency metrics that fail to adequately address misalignments in process, intent, or social compatibility. The goal for improving AI agents should be to

[5] Following the same logic as with single points of failures and technological monoculture risks, the design of or changes in the AI agent substrate (e.g., the underlying foundation model or LLM) can change the agentic behavior for large numbers of users. This is a form of dependency and influenceability that is different in scale (one-to-many, i.e., the same AI model is the basis for agents used by large numbers of individual users), timeliness, and effectiveness compared to human agent scenarios.

[6] While arguably important, note that we do not discuss privacy or security issues in this essay.

[7] See, for example, Batzner et al. (2025a, 2025b), Pan et al. (2025), and De Freitas et al. (2025).

enhance human goals, not to make AI systems more valuable.[8] If the interests and goals of all principals were perfectly aligned, the optimal design of agents might be viewed as a purely technical challenge, but that is not the case. The range of causes for (value) misalignments across different stakeholder groups in PAP contexts persists in AI systems (e.g., Gabriel et al., 2024 and 2025; Sorensen et al., 2024; Whitt, 2024b).

Moreover, attempts to enhance AI agents can render the system more complex, paradoxically challenging transparency goals. For example, it is plausible that the very technological basis that enables GenAI's capabilities renders them inherently complex to steer. Examples are black box characteristics and non-predictable outcomes. While model steerability may be limited and subject to what we could call 'model inertia', closed commercial AI models typically restrict user access to parameter modifications, limiting customization options compared to (more) open model alternatives (e.g., Schrepel and Pentland, 2024). For comprehensive overviews of how different (pre- and post-) training approaches can shape model behavior and outcomes, see, for example, Ohm (2024) and Tie et al. (2025).

Finally, misalignment may naturally result if customization needs are unknown and/or unaddressed.[9] PAP-related optimization needs are local, dispersed contextually (e.g., with respect to problem domain, task, principal identity, etc.), and dynamic (e.g., agents adapt and learn from interactions and become dynamically personalized), which causes control challenges that can only be solved if an appropriate measurement ecosystem is available to ensure the ability to detect, protect, and act.

While some AI agent developers may design, develop (and align), and deploy AI systems customized to their needs, even using their own data, other developers/users may rely more heavily on pre-trained models and use post-training to customize them. Entities that design or deploy AI models can shape those based on various pre-training and post-training approaches that differ significantly in terms of effectiveness and resource, data and skill requirements. Examples include the inability of individual users to customize an AI system beyond simple interventions, such as prompt engineering or in-context learning. When users rely on plain-vanilla, out-of-the-box commercial AI agents or those provided by consumer-facing platforms, power hierarchies and the directionality of influence between principals and agents may be inverted. For example, Batzner et al. (2025a) offered a discussion of sycophantic and similar behavior in LLMs. A good agent should not try to befriend users with flattery but provide advice that advances the users' best interests. Less

[8] Using the words of Ji et al. (2025, p. 4), AI alignment focuses on processes and approaches "to make AI systems behave in line with human intentions and values" (Ji et al., 2025, p. 4), with alignment approaches "focusing more on the objectives of AI systems than their capabilities."

[9] Shur-Ofry (2025, Abstract) cautioned of "the propensity of large language models (LLMs) to generate mainstream, standardized contents, potentially narrowing their users' worldviews."

sophisticated users remain more vulnerable to agentic influence that may to a greater or lesser extent be aligned with their individual interests and may imply that they are nudged in ways more consistent with 'shadow principals' rather than in the end-user's interests. The more technologically sophisticated users are in terms of their skills and literacy, and the more resources and data they have, the more they will be able to build and develop customized AI systems. These users will likely be in a better position to counter and mitigate such dependencies.

Furthermore, the deeper and more persistent human-AI relationships become, the more critical are feedback loops that shape mutual influence between humans and AI systems. Influence is bidirectional, and human goals, preferences, and values shape – and are shaped by – adaptive, personalized AI systems.[10] This interdependence highlights a co-evolution that renders alignment challenges in the PAP context inherently dynamic and complex.

Finally, and paradoxically, the better an AI agent serves human interests, the greater the user's trust and dependency – and vulnerability if the agent turns rogue.

3. Digital (Platform) Economics on Agentic AI Steroids

As the preceding makes clear, the PAP framework in which AI agents will operate will be multi-agent systems in which agents – on behalf of humans, other entities, and among each other – can bargain and transact. At its core, agentic AI creates novel opportunities for distributed AI, empowering various ecosystem stakeholders through facilitating more granular, individualized, and personalized decision-making and direct transactions – at scale and in real-time. This fundamentally challenges the conventional role and value (proposition) of digital platforms as aggregators, intermediaries that manage network effects within and across interdependent market sides (i.e., user groups such as buyers, sellers, and advertisers), transaction facilitators and trust builders, and so on.

Although we see significant potential for widespread disruptions, the effects and directions of the anticipated changes will vary significantly across different market contexts,[11] and remain impossible to reliably forecast. For example, we should expect that powerful ecosystem actors like large tech (platform) companies will adapt and lead this change, seeking novel avenues to maintain or even expand their role in AI-driven digital economies, but figuring out which stakeholders will be most successful is uncertain.[12]

---

[10] See, for example, Shen et al. (2025).

[11] For example, durable versus consumable goods, search versus experience goods, mass-market versus niche goods, etcetera.

[12] The uncertainty prevails in spite of there being no shortage of forecasters; and undoubtedly some forecasts will prove better (ex post) than others. See, in this context, also the discussion by Whitt (2024b)

How might a future of direct agentic AI bargaining transform digital platform economics? It is certainly possible that AI-agent-to-AI-agent interactions will significantly reduce the transaction costs associated with identifying and consummating matching transactions, benefiting both parties. It is also possible that the commerce between consumers, digital platforms, and service/product providers (on the other side of, say, a two-sided platform) will be changed by systemic shifts in bargaining power. For example, if the buy-bots of consumers and sell-bots of vendors (or, bots of the digital platforms that match consumers and vendors) all interact, whose bots would we expect to be the better bots? And if there are benefits from better bots or agents, will this give rise to an arms race favoring the privileged who have more access to resources such as compute, data, labor, and skills – thereby reinforcing power imbalances within digital economies?[13]

In a world of multi-agent ecosystems and agentic automation, the potential for bots to reduce transactions to customized "market of one" transactions may be increased. If so, that might shift outcomes from a neoclassical competitive marketplace of many independent sellers interacting with many independent consumers, with supply and demand being balanced by an emergent market price, to bilateral bargaining with the price being the result of negotiations. These negotiations are shaped by bargaining power and that can fail or increase transaction costs in the form of deadweight investments in bargaining position.[14]

Because the effects will be highly market-context dependent and outcomes are infeasible to forecast reliably, effective AI governance will rely on agile adaptive frameworks and guardrails that can only be adequately designed and effectively enforced if we have sufficient information and understanding about agentic bargaining. A key requisite will be a measurement ecosystem capable of evaluating these transactions and bargaining outcomes (see also Lehr and Stocker, 2024; Stocker and Lehr, 2025).

The impact of Agentic AI may be especially profound for the economics of advertising supported digital platforms. Agentic mediation breaks the link between human eyeball attention and content on websites and other digital platforms. Instead

about edge tech to counter the power by large tech firms and Kapoor et al. (2025)'s discussion of platform agents that could reinforce and entrench incumbent platform power while other agents controlled by users, seek to safeguard user autonomy.

[13] Zhu et al. (2025) present the results of a simulation experiment of LLM agent mediated e-commerce, concluding that their "analysis reveals that agent-to-agent negotiation is naturally an imbalanced game where users with less capable agents face significant financial loss against stronger agents" (Zhu et al., 2025, p. 8). See also Chen (2025).

[14] One extreme form of bargaining asymmetry tilted to the producer/seller side would result in first degree price discrimination where consumer surplus is fully extracted. Alternatively, the fragmentation of markets might foreclose some choices, denied sufficient scale and liquidity for economic viability.

of human users consuming content and browsing through products on marketplaces, agents do that on behalf of users. Agents may also negotiate on behalf of users. This is at odds with traditional online business models predicated on a direct flow of human user (aka "eyeball") attention to content. Traditional network effects often relied on direct user engagement and attention – if humans no longer visit sites directly, 'traditional' human-user-based network effects may weaken.

Today, advertisers fund much of the web, but they pay for human exposure and are incentivized to invest by a platform's ability to attract human attention. When agents replace or augment humans as 'primary consumers', the role of platform intermediaries and related digital economics changes. First, conventional metrics for measuring engagement and reach must be reimagined as the nature and locus of user interaction change. When bots controlling bots are the source of website impressions and click counts, today's key currency of our digital economy is instantly devalued. Second, the locus of monetization channels may change. Companies may redirect their advertising budgets from websites and other consumer-facing platforms directly to AI development, helping to establish new avenues for (covert or shadow principal) influence.[15] Incentives and efforts to explore novel technological possibilities for influencing/manipulating agent outputs to extract (more) consumer surplus can be expected to grow.[16] Third, while memory and personalization via adaptive and self-evolving agents may offer novel ways to lock in customers, AI agents like buy bots may empower consumers by enabling them to identify and exploit arbitrage opportunities and switch between platforms. However, this can only work if agents have access to the relevant information, and it also emphasizes that the need for discoverability – a key affordance of digital platforms – may be achieved through more mechanical databases that list offers and can be used by AI agents or crawlers (e.g., via APIs).[17]

These changing digital (platform) economics imply a strong potential for redistributing market power (e.g., due to shifts in bargaining costs or control over and influence of agents). Controlling a complex system that we understand is difficult; controlling an agentic AI system in which we may not know whose agent is

[15] See Bradley (2025), Guevara (2025), Kumar and Lakkaraju (2024), or Sommerfeld et al. (2025) for recent research on how (Gen)AI/LLMs may alter the economics of advertising supported platforms, with implications for digital tools and strategies like SEO and the control and targeting of advertising.

[16] See Allouah et al. (2025) for research on how vision language models may enhance advertising effectiveness. More generally, agents may shift their focus to anticipate the behavior of other agents rather than the human principals whose interests they were originally launched to represent. There are numerous examples of how machine-to-machine transactions can lead to dysfunctional market outcomes, including the program-trading interactions that resulted in the stock market crash in 1987 and the algorithmic pricing that lead to a book about flies being priced at $24 million (see Solon, 2011).

[17] Moreover, GenAI can be expected to exacerbate matching problems by facilitating the potential for the artificial generation of content of all sorts (e.g., Cooper, 2025; Cooper et al., 2025), causing growing gluts of content in research papers, articles, and media at scale, and thereby increasing the need for AI agents to assist in sorting through all of the content (i.e., who – or what – will read it all?).

interacting with whose agent, what machine-to-machine linkages may be shaping their interactions, and how differences in agent capabilities shape bargaining outcomes will be significantly more challenging.

4. Conclusion

What is clear from this brief exploration of PAP dynamics and digital (platform) economics in the context of an agentic AI future is that a host of complex challenges loom. The shift from direct user-platform interactions to agent-mediated relationships introduces complex principal-agent dynamics and new levers and avenues for exercising control and influence over AI systems. Two obvious implications are the potential to manipulate information and users, as well as to disrupt existing business models that rely on user attention.

Understanding these transformations requires careful attention to both the technical design and capabilities of AI agents, as well as the economic and social structures within which they operate. As AI technologies continue to evolve, the interplay between user control, agent autonomy, and platform power will shape not only PAPs and technical and economic alignment problems but also determine the ultimate impact on digital markets and social welfare.

The challenge moving forward will be to harness the efficiency gains from AI agents while maintaining user agency, preventing manipulation, and ensuring that the benefits of these technologies are broadly distributed rather than concentrated among a few powerful ecosystem actors. This is easier said than done. As AI is in many ways becoming more distributed and personalized, effective and meaningful AI governance must be predicated on a workable multi-stakeholder measurement ecosystem that maintains the ability to detect, protect, and act – even in a world where measurement challenges are becoming increasingly complex and dynamic and measurement needs are becoming more fragmented, dispersed, and granular.

Volker Stocker & William Lehr

This work benefited greatly from inspiring conversations with and very helpful feedback from Zachary Cooper and Jan Batzner. Volker Stocker would like to acknowledge funding by the Federal Ministry of Education and Research of Germany (BMBF) under grant No. 16DII131 (Weizenbaum-Institut für die vernetzte Gesellschaft – Das Deutsche Internet-Institut)

References

1. Allouah, A., Besbes, O., Figueroa, J. D., Kanoria, Y., & Kumar, A. (2025). What Is Your AI Agent Buying? Evaluation, Implications and Emerging Questions for Agentic E-Commerce. arXiv preprint arXiv:2508.02630.

2. Batzner, J., Stocker, V., Schmid, S., & Kasneci, G. (2025a). Sycophancy Claims about Language Models: The Missing Human-in-the-Loop. ICLR 2025 Workshop on Bidirectional Human-AI Alignment.
3. Batzner, J., Stocker, V., Tang, B., Natarajan, A., Chen, Q., Schmid, S., and Kasneci, G. (2025b). Whose Personae? Synthetic Persona Experiments in LLM Research and Pathways to Transparency. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society.
4. Bolton, P. and Dewatripont, M. (2005). Contract theory, MIT Press: Cambridge MA.
5. Bradley, S. (2025). After a year of Google’s AI Overviews, marketers consider tweaking their paid search strategies. Digiday, 16 May. Available at: https://digiday.com/marketing/after-a-year-of-googles-ai-overviews-marketers-consider-tweaking-their-paid-search-strategies/
6. Capgemini (2025). Rise of Agentic AI. 15 July. Available at: https://www.capgemini.com/wp-content/uploads/2025/07/Final-Web-Version-Report-AI-Agents.pdf
7. Caserman, A. (2025). 2025 Is the Year of Agentic AI. Medium, 3 March. Available at: https://medium.com/@alenka.caserman/2025-is-the-year-of-agentic-ai-5aa3bfd8b8c6
8. Chen, C. (2025). When AIs bargain, a less advanced agent could cost you. MIT Technology Review, 17 June. Available at: https://www.technologyreview.com/2025/06/17/1118910/ai-price-negotiation/ .
9. Cooper, Z. (2025). Dams for the Infinite River: Limits to Copyright’s Power over the Next Generation of Generative AI Media. Network Law Review. Available at: https://www.networklawreview.org/cooper-gen-ai/
10. Cooper, Z., Lehr, W.H., & Stocker, V. (2025). The New Age: Legal & Economic Challenges to Copyright and Creative Economies in the Era of Generative AI. The Digital Constitutionalist, 23 January. Available at: https://digi-con.org/the-new-age-legal-economic-challenges-to-copyright-and-creative-economies-in-the-era-of-generative-ai/
11. De Freitas, J., Oğuz-Uğuralp, Z., & Kaan-Uğuralp, A. (2025). Emotional Manipulation by AI Companions. arXiv preprint arXiv:2508.19258.
12. Deslandes, N. (2025). 2025 Informed: the Year of Agentic AI. Tech Informed, 11 January. Available at https://techinformed.com/2025-informed-the-year-of-agentic-ai/
13. Gabriel, I., Manzini, A., Keeling, G., Hendricks, L. A., Rieser, V., Iqbal, H., ... & Manyika, J. (2024). The ethics of advanced ai assistants. arXiv preprint arXiv:2404.16244.

14. Gabriel, I., Keeling, G., Manzini, A., & Evans, J. (2025). We need a new ethics for a world of AI agents. Nature, COMMENT, 4 August. Available at: https://www.nature.com/articles/d41586-025-02454-5
15. Gao, H. A., Geng, J., Hua, W., Hu, M., Juan, X., Liu, H., ... & Wang, M. (2025). A survey of self-evolving agents: On path to artificial super intelligence.arXiv preprint arXiv:2507.21046.
16. Guevara, W. (2025). Google AI Overviews: New CTR Study Reveals How to Navigate Negative SERP Impact, AMSIVE, 16 April. Available at: https://www.amsive.com/insights/seo/google-ai-overviews-new-research-reveals-how-to-navigate-click-drop-off/
17. Jarrahi, M.H., & Ritala, P. (2025). Rethinking AI Agents: A Principal-Agent Perspective. California Management Review: CMR Insights, 23 July. Available at: https://cmr.berkeley.edu/2025/07/rethinking-ai-agents-a-principal-agent-perspective/
18. Jensen, M.C., and Meckling, W.H. (1976). Theory of the firm: Managerial behavior, agency costs and ownership structure. Journal of Financial Economics, 3(4), 305-360.
19. Ji, J., Qiu, T., Chen, B., Zhang, B., Lou, H., Wang, K., ... & Gao, W. (2025). AI alignment: A comprehensive survey. V6. arXiv preprint arXiv:2310.19852v6.
20. Kapoor, S., Stroebl, B., Siegel, Z. S., Nadgir, N., & Narayanan, A. (2024). AI agents that matter.arXiv preprint arXiv:2407.01502.
21. Kapoor, S., Kolt, N., & Lazar, S. (2025). Build Agent Advocates, Not Platform Agents. arXiv preprint arXiv:2505.04345.
22. Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI agents for alignment and governance.arXiv preprint arXiv:2504.21848.
23. Kirk, H. R., Gabriel, I., Summerfield, C., Vidgen, B., & Hale, S. A. (2025). Why human–AI relationships need socioaffective alignment. Humanities and Social Sciences Communications. 12(1), 728. doi:10.1057/s41599-025-04532-5
24. Kolt, N., 2025. Governing AI agents. Notre Dame Law Review, Vol. 101, Forthcoming. Available at: http://dx.doi.org/10.2139/ssrn.4772956
25. Kumar, A., & Lakkaraju, H. (2024). Manipulating large language models to increase product visibility.arXiv preprint arXiv:2404.07981.
26. Laffont, J.J., & Martimort, D. (2002). The Theory of Incentives: The Principal-Agent Model. Princeton University Press. https://doi.org/10.2307/j.ctv7h0rwr.
27. Lehr, W.H., & Stocker, V. (2024). Competition Policy over the Generative AI Waterfall. In: A. Abbott & T. Schrepel (eds.) AI and Competition Policy, Concurrences (pp. 335-358).

28. Ménard, C., & Shirley, M. M. (2022). Advanced Introduction to New Institutional Economics. Edward Elgar Publishing. https://doi.org/10.4337/9781789904499
29. Milgrom, P., & Roberts, J. (1992). Economics, organization and management (1st ed.). Prentice Hall.
30. Ohm, P. (2024). Focusing On Fine-Tuning: Understanding The Four Pathways For Shaping Generative AI. Science and Technology Law Review, 25(2). https://doi.org/10.52214/stlr.v25i2.12762
31. Pan, X., Fan, J., Xiong, Z., Hahami, E., Overwiening, J., & Xie, Z. (2025). User-Assistant Bias in LLMs.arXiv preprint arXiv:2508.15815.
32. Qian, C., Zhu, K., Horton, J., Manning, B. S., Tsai, V., Wexler, J., & Thain, N. (2025). Strategic Tradeoffs Between Humans and AI in Multi-Agent Bargaining.arXiv preprint arXiv:2509.09071v2. https://doi.org/10.48550/arXiv.2509.09071
33. Schneier, B. (2025). AI and Trust.Communications of the ACM, 68(8), 29-33.
34. Schrepel, T. and Pentland, A. (2024). Competition between AI foundation models: dynamics and policy recommendations. Industrial and Corporate Change, 2024, dtae042. https://doi.org/10.1093/icc/dtae042
35. Shaw, F.X. (2025). Microsoft Build 2025: the Age of AI Agents and building the open agentic web. Microsoft Blog, 19 May. Available at: https://blogs.microsoft.com/blog/2025/05/19/microsoft-build-2025-the-age-of-ai-agents-and-building-the-open-agentic-web/
36. Shen, H., Knearem, T., Ghosh, R., Liu, M. X., Monroy-Hernández, A., Wu, T., ... & Hearst, M. (2025). Bidirectional Human-AI Alignment: Emerging Challenges and Opportunities. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (pp. 1-6). https://doi.org/10.1145/3706599.3716291
37. Shur-Ofry, M. (2025). Multiplicity as an AI governance principle.Indiana Law Journal, 100. https://www.repository.law.indiana.edu/ilj/vol100/iss4/6
38. Solon, O. (2011). How A Book About Flies Came To Be Priced $24 Million On Amazon. WIRED, 27 April. Available at: https://www.wired.com/2011/04/amazon-flies-24-million/
39. Sommerfeld, N., Dave, R., and Webster-Clark, D. (2025). Marketing’s New Middleman: AI Agents, Bain & Company Brief. Available at: https://www.bain.com/insights/marketings-new-middleman-ai-agents/
40. Sorensen, T., Moore, J., Fisher, J., Gordon, M., Mireshghallah, N., Rytting, C.M., Ye, A., Jiang, L., Lu, X., Dziri, N., Althoff, T., and Choi, Y. (2024). Position: a roadmap to pluralistic alignment. In Proceedings of the 41st

International Conference on Machine Learning (ICML'24), Art. 1882, 46280–46302.

41. Stocker, V., & Lehr, W.H. (2025). The Growing Complexity of Digital Economies over the GenAI Waterfall: Challenges and Policy Implications. Network Law Review. Available at: https://www.networklawreview.org/stocker-lehr-ecosystem/
42. Tie, G., Z. Zhao, D. Song, F. Wei, R. Zhou, Y. Dai, W. Yin, Z. Yang, J. Yan, and Y. Su (2025), A survey on post-training of large language models. https://arxiv.org/abs/2503.06072v3
43. Whitt, R. (2024a). Rise of the KnowMeBots: Promoting the Two Dimensions of AI Agency.Colorado Technology Law Journal, 23 (1) 49.
44. Whitt, R. (2024b). Reweaving the Web. How together we can create a human-centered Internet of trust.
45. Zhu, S., Sun, J., Nian, Y., South, T., Pentland, A., & Pei, J. (2025). The Automated but Risky Game: Modeling Agent-to-Agent Negotiations and Transactions in Consumer Markets. V4. arXiv preprint arXiv:2506.00073v4.