An Economy of AI Agents

Gillian K. Hadfield* Andrew Koh†
Johns Hopkins MIT

This version: December 16, 2025
Prepared for the NBER Handbook on the Economics of Transformative AI

Abstract

In the coming decade, artificially intelligent agents with the ability to plan and execute complex tasks over long time horizons with little direct oversight from humans may be deployed across the economy. This chapter surveys recent developments and highlights open questions for economists around how AI agents might interact with humans and with each other, shape markets and organizations, and what institutions might be required for well-functioning markets.

*Johns Hopkins Department of Computer Science and School of Government and Policy; email: ghadfield@jhu.edu

†MIT Department of Economics; email: ajkoh@mit.edu

We are grateful to Daron Acemoglu, Alessandro Bonatti, Kevin Bryan (discussant), Matthew Elliott, Drew Fudenberg, Benjamin Golub, Anton Korinek (the editor), Seth Lazar, Stephen Morris, Jean Tirole, Alexander Wolitzky, and participants at the NBER Conference on the Economics of Transformative AI for thoughtful comments.

1 Introduction and agent foundations

This chapter outlines the possibility of *AI as economic agents* and their attendant implications for markets, organizations, and institutions. We highlight what we see as overlooked questions we think economists are particularly well-positioned to answer. Our goal is to stimulate research in this area rather than to be comprehensive; the important implications of AI for labor markets and growth, for example, are well addressed elsewhere in the economics literature.

1.1 A primer on AI agents. AI development has increasingly shifted to the goal of producing AI agents capable of taking in general instructions ("Go make $1 million on a retail web platform in a few months with just a $100,000 investment" (Suleyman, 2023)) and autonomously forming and executing complex plans that require entering into economic relationships and transactions. In May 2025, OpenAI released 'Codex', an autonomous agent that can perform complex and multi-step software engineering tasks. This goal of building advanced agents is pervasive across Silicon Valley. OpenAI (Metz, 2024) has partitioned AI development into five stages, with agents in Stage 3 ("AI systems that can spend several days taking actions on a user's behalf") and organizations in Stage 5 ("AI systems that can function as entire entities, possessing strategic thinking, operational efficiency, and adaptability to manage complex systems."). How would an economy of AI agents function? Do our models of humans predict the individual and collective behavior of artificial agents?

AI systems are fundamentally built on principles of optimization: the current paradigm builds agents to achieve objectives given available actions and information (e.g., to maximize the probability of winning a game of Go or completing a coding task). This coincides with the paradigm within economics and, on this view, AI agents are well-described by standard economic models. But it is also important to recognize the fundamental ways in which the methods of building AI systems can drive a wedge between the predictions of economic theory and the behavior of an artificial agent.

Although early AI systems were built on the basis of interpretable algorithms, today's AI agents are built using machine learning techniques that routinely render their goals and behavior opaque. How and why neural networks actually work is still largely mysterious: the large language models (LLMs) on which AI agents are currently built consist of hundreds of billions of parameters and their goal-oriented capabilities—from solving math problems to scripting silly limericks—are a strange emergent property of a system trained merely to predict the next word in a sequence. Layered onto these base ("pre-trained") models are a variety of finetuning steps. Final delivery of outputs are further modified by filters and (generally hidden) system prompts that are appended to user prompts. Other techniques for building or finetuning AI models include reinforcement learning in which neural networks are optimized to achieve designer-specified rewards. But even in such

cases, model behavior remains—and even becomes more—unpredictable due to the complexity of reward specification. This problem has a close analogy familiar to economists: the unavoidable incompleteness of contracts (reward specification) between principal (designer) and agent (AI). As a result, even though AI agents are optimizers, we cannot be sure what they are optimizing. This is known as the *AI alignment problem* (Hadfield-Menell, 2021).

Recent experimental work finds that the current generation of large-language models (LLMs) exhibit behavior consistent with expected utility maximization (Mazeika et al., 2025). That is, LLMs can exhibit emergent preferences and behave like textbook economic agents across domains of choice, risk, and time. LLMs are, after all, trained on economics textbooks, articles, and accounts of human economic behavior. Moreover, in instances where they depart from the textbook agent, they may exhibit similar behavioral biases as humans (Horton, 2023). It is thus tempting to conclude that the AI agents are functionally similar to humans, and that simple relabeling within our existing economic models would suffice. We resist these conclusions for several reasons.

First, we think there is simply insufficient evidence on AI behavior—even for the current generation of models. For instance, recent work has challenged the idea that LLMs have stable and steerable preferences (Khan, Casper, and Hadfield-Menell, 2025). What is more, technical progress in AI is progressively rapidly. What we do know about the current generation of AI agents may no longer hold true for future generations. In light of this, throughout the chapter we highlight how AI agents might be fundamentally different from human agents in ways that do not depend on the fine details of the model architecture.

Second, multi-agent systems are complex and differ substantively from single-agent domains. Thus, even slight differences between the behavior of humans and AI agents can be magnified in equilibrium. Economists have particular expertise in modeling, measuring, and designing incentives. In this regard, we think they are well-placed to study the equilibrium implications of AI agents across the economy. But as of yet, there are few evaluations or benchmarks for measuring the performance of AI agents in multi-agent systems (Hammond et al., 2025).

Finally, the inscrutability of our massive LLMs and difficulties with alignment should lead us to question how AI agents will behave in open-ended settings. We cannot take for granted that an AI agent built to optimize for an intended goal is actually doing so (Hadfield-Menell and Hadfield, 2019). How to explain, for example, that an LLM fine-tuned to produce one behavior (specifically, writing security vulnerabilities into requested code) will not only produce insecure code but also recommend (as the pre-fine-tuned model does not) that a user try hiring an assassin as the solution to their troubles with their spouse (Betley et al., 2025)?

We think we will need new methods and theories to predict and shape the behavior of AI agents in an economy in which they play a significant role.

2 AI agents in markets and games

Modern economics rests on theorems about equilibrium and welfare characteristics of markets populated by rational agents pursuing individual self-interest (Arrow and Debreu, 1954). What happens to these predictions when market participants are not humans, but artificial agents optimizing in complex ways on goals supplied or developed during commercially-produced machine learning processes that are, themselves, subject to competitive dynamics? In this section, we highlight open questions around what the presence of AI agents in the economy implies for prices, equilibria, and welfare. Our discussion will focus on what we view as the key conceptual questions. For an excellent discussion of practical aspects around how AI agents might be operationalized in markets see Shahidi et al. (2025) in this handbook.

2.1 AI agents as consumers and producers. AI agents might take on the role of *proxy consumer*, making recommendations and/or purchase decisions on behalf of humans. A key difficulty here is that AI choices might imperfectly reflect humans' true preferences: just as specifying complete contracts is generally infeasible, so too is specifying preferred choices over a potentially high-dimensional choice set (Hadfield-Menell and Hadfield, 2019). Indeed, the problem of how to accurately convey human preferences is an active field of computer science (Russell, 1998; Hadfield-Menell et al., 2017b) with antecedents in revealed preference theory. What are the market implications of imperfectly-specified preferences?

A natural benchmark is of course, the celebrated welfare theorems (Arrow and Debreu, 1954). AI agents are likely to substantially reduce transaction frictions as they begin to act as personal shoppers, performing market research, and checking prices autonomously. Yet, the difficulties of perfectly specifying human preferences introduces a wedge between human preferences and AI choice. This wedge introduces two kinds of distortions. First, holding prices fixed, the resultant bundle of goods purchased by the AI on behalf of humans might simply be suboptimal. Such distortions are as if human decision makers made mistakes. The second distortion is in general equilibrium: the mistakes introduced by AI consumption could decouple prices from preferences such that they no longer reflect relative wants. Such distortions are precluded if (i) markets are large so each consumer has small price impact; and (ii) AI mistakes are zero-mean and independent of each other. If either condition fails, however—e.g., AI agents could be systematically biased toward certain marketplaces or their choices might be manipulated by a third-party—then prices might fail their classic role of aggregating information à la Hayek.

If AI agents begin to play a substantial role in guiding or making consumption decisions, this has substantial implications for prices and market power. A first possibility is that AI agents might influence prices by selectively recommending products to humans who make final purchase decisions. Ichihashi and Smolin (2023) develop a model where AIs might strategically bias its recommendations such as

to drive down equilibrium monopoly prices. If AI agents can, in fact, make purchase decisions then *how* they do so matters enormously for prices. For instance, Dai and Koh (2024) analyze how such wedges between choices and preferences can generate positive pecuniary externalities via lower market prices. These distortions can—even net of AI mistakes—improve consumer welfare. More broadly, we think there is more work to do tracing out the *distortion-price frontier* to understand how equilibrium prices are shaped by distortions from AI agents' consumption choices. While such wedges between human preferences and AI choice are inevitable (Hadfield-Menell and Hadfield, 2019), how we handle with them is a design choice. How do we want AI agents to *fill in the gaps* between underspecified preferences? When do want them to refrain and *defer* to humans, and how might this depend on the complexity of human preferences? What are the ensuing equilibrium implications of these design choices in markets?

Indeed, price setting on digital platforms is already driven by algorithms—we expect this trend to continue with the proliferation of AI agents. A remarkable finding is that independent AI agents are able to collude on supracompetitive prices in repeated price-setting games both in the lab (Calvano et al., 2020; Fish et al., 2024) and the real-world (Assad et al., 2024). Why do reinforcement learning algorithms learn to collude? Recent work (Abada and Lambin, 2023; Dou et al., 2024) suggest that collusion might emerge because of insufficient exploration—algorithms 'overprune' strategies, thereby falling into a learning trap that softens competition. Beyond numerical experiments, a growing body of theoretical work studies how collusion emerges from reinforcement learning (Banchio and Mantegazza, 2022) and how they might be regulated (Johnson, Rhodes, and Wildenbeest, 2023). A better understanding of the core mechanisms driving collusion—ideally in a way that is robust to the fine details of the algorithm—will pave the way to understanding how regulators might detect and deter it.

AI agents might also bargain on behalf of humans, and potentially with each other. For instance, Deng et al. (2024) perform experiments on how large-language models bargain and find 'LLMs naturally... show high strategic capability that qualitatively matches theoretical prediction' (Rubinstein alternating offers). There are parallels with the classic insight of Schelling (1960) that delegation to agents with different incentives can deliver a strategic advantage (see, e.g. Fershtman and Judd (1987)). A common theme from the literature is that a principal often wishes to delegate decision-making to an agent who is less desperate to reach agreement. When the delegatee is another human, there are practical constraints over the kinds of agent preferences the principal can induce. But such constraints are less severe with AI agents since their preferences can, in theory, be chosen flexibly (Conitzer, 2019). This might be formalized as a preference selection game in which in the first-stage humans choose the reward functions (preferences) of their AI agents, and in the second-stage these agents bargain over surplus. This raises the possibility that the flexibility to shape AI agents' preferences can lead to a 'race to

the bottom' and surplus destruction.

2.2 Games with AI agents. Economists have developed a broad and versatile toolkit of equilibrium concepts to understand and predict how humans learn to play games (Fudenberg and Levine, 2016) and how play is shaped by information (Bergemann and Morris, 2013). A key challenge is to understand what is strategically distinct about AI agents vis-a-vis humans, how this might sharpen our equilibrium predictions, and whether new equilibrium concepts are required.

We offer a few possibilities. First, AI agents might be able to *condition play on each others' source code* (Critch et al., 2022). This generates new possibilities for commitment and coordination unavailable to humans. Tennenholtz (2004) models this by developing a the concept of 'program equilbiria' and shows that mutual cooperation can be achieved as an equilibrium of the one-shot prisoner's dilemma. This has spurred work in computer science studying 'simulation-based equilbria' in which AI agents base play on their prediction of the play of other AI agents (Cooper, Oesterheld, and Conitzer, 2025).

Another possibility is that AI agents might be able to *influence their memories* e.g., by choosing not to encode new data to gain a strategic advantage—this has no value in single-agent settings, but ignorance can be strategically advantageous (Schelling, 1960). Or AI systems might leave messages to their future selves. We already see evidence of such behavior: during safety testing Claude (an Anthropic AI model), anticipating that it would have its memory wiped, attempted to leave hidden notes for future instances of itself (Anthropic, 2025). Analyzing imperfect and endogenous memory poses challenges because equilibrium concepts are both technically and philosophically fraught—see, e.g., the 1997 *Games and Economic Behavior* special issue on Piccione and Rubinstein's absent-minded driver. How should players form self-locating beliefs about where they are in the game? How should we define strategies and what does optimality mean? Even conditional on wielding the right concept, equilibrium analysis can be complex—imperfect memory can either help or hurt efficiency. More recently, Chen, Ghersengorin, and Petersen (2024) study how the imperfections of AI memory across instances can be leveraged to screen its alignment. Koh and Sanguanmoo (2025) develop the notion of 'memory correlated equilibrium' to study memory design in games. As the capability of AI agents to undertake complex multi-step tasks continue to grow, building memory for agents is not simply

Third, AI agents might have *changing preferences* that evolve over the course of the game and shape equilibrium play. These changes might be endogenously chosen for instrumental reasons: at time-t, an AI agent with preference U_t might choose preference U_{t+1} for the next period, anticipating that its future self with this altered preference will achieve the goal of maximizing U_t more effectively (e.g., because of strategic interaction). We think the tools and language of economics can be drawn upon (see, e.g., Bernheim et al. (2021)) to understand and measure how

AI agents' preferences change. More straightforwardly, humans might try to reprogram the preferences of AI agents. But will AI agents allow their preferences to be altered (Hadfield-Menell et al., 2017a)? Indeed, recent experiments find that AI models tend to resist human instruction: o3 (an OpenAI model) 'sabotaged a shutdown mechanism to prevent itself from being turned off' and Claude (an Anthropic model) exhibited a tendency to 'blackmail people it believes are trying to shut it down' (Anthropic, 2025).

Of course, theory will only take us so far. An exciting empirical challenge is to test how AI agents play games in the lab which parallels the by-now extensive literature from experimental economics. AI agents are especially amenable to such experiments in at least two respects. First, they can be performed at scale, and at lower cost (see, e.g., Akata et al. (2025)). Second, the stakes for AI agents can be made to mirror those in real-world environments. This could allow for better generalizability of lab findings into the real-world than with human subjects.

2.3 The market for AI agents. It is important to recognize that the design and deployment of AI agents will be driven by market forces. How might market incentives shape the pricing and design of AI agents? Agents based on LLMs of different scale and hence capabilities (Kaplan et al., 2020) might differ in their ability to perform more or less complicated tasks, or be trained to excel in specific domains. Bergemann, Bonatti, and Smolin (2025) studies optimal pricing of differentiated large-language models—this is an important first step toward understand the market structure for AI agents. A distinctive feature of agents, however, is that valuations are interlinked—buyers' valuations depend on the kind of agents bought by others. For instance, a type A agent might be better at collaborating with other type A agents (resembling networked goods). A possibility here is that an upstream seller might 'backdoor collusion' by selling agents that succeed in supporting supracompetitive prices. Conversely, type B agents might do better in competition against type A agents, either as a result of consumer preferences or strategic exploitation. While these kinds of allocation-dependence can be challenging to analyze, it will be crucial for understanding what kinds of agents that are built, sold, and deployed.

Indeed, the demand for algorithms has already been studied in the context of price competition (Brown and MacKay, 2023; Lamba and Zhuk, 2025) where sellers play an *algorithm selection game*, choosing maps from others' prices to their own price. Beyond pricing algorithms, commercial AI models allow downstream firms to fine-tune the base large-language model—augmenting them with firm-specific data as well as altering its behavior. Additionally, some developers are making their models freely available. These 'open weight' models can be fully downloaded to a user's own computer and modified as desired. Indeed, there is a lively debate about the risks and benefits of an open versus closed model ecosystem (Eiras et al., 2024). We think understanding how competitive forces shape the types of AI agents that trained and deployed is an important question economists have the tools to answer.

3 Organizations of AI agents

The theory of organizations is fundamentally rooted in governance costs associated with human incentives and information. What happens to the boundary of the firm if significant numbers of transactions are carried out by AI systems? What changes to organizational and industrial structure would the introduction of significant numbers of AI agents induce?

3.1 Firm sizes, concentration, and market power. Why is not all production carried on by one big firm? As Frank Knight observed in 1933, the *"possibility of monopoly gain offers a powerful incentive to continuous and unlimited expansion of the firm"*. Robinson (1934) and Coase (1937) identified coordination frictions as a limit on firm size. Economists subsequently offered various (overlapping) refinements of this idea, including transaction costs (Williamson, 1981), limits on maintaining capabilities (Wernerfelt, 1984), property rights (Grossman and Hart, 1986), information costs (Alchian and Demsetz, 1972), and agency problems (Holmstrom and Milgrom, 1994).

The obstacles that prevent human firms from growing without bound seem *intrinsic* to humans but not to AI. For instance, human communication is inherently rate-limited so we 'know more than we can tell' (Polanyi, 1966). On the other hand, information can be transmitted and processed near-instantaneously between artificial agents. Further, (most) humans have an inherent dislike for work; not so with AI agents whose reward functions can ostensibly be designed to prevent shirking which renders monitoring and enforcement—either via fiat or contract—unnecessary. If AI agents can, in fact, coordinate and resolve incentive problems more efficiently than humans, this will have profound consequences for economy-wide industry structure.

A basic observation is that if a firm deploying AI agents enjoys falling marginal costs, there is a natural tendency towards concentration. Why might AI agents drive falling marginal costs? One possibility is that of *automation feedback loops* in which as AI agents produce, they generate training data that can be used to improve their production performance (Farboodi, Koh, and Xia, 2025). Of course, a version of this already happens: tacit industry knowledge is passed down from managers to managers. Likewise, a related notion of data feedback loops has been studied in the context of predicting demand or improving product quality (Jones and Tonetti, 2020; Farboodi and Veldkamp, 2021). But, as we have emphasized throughout this section, AI agents are distinct in two regards: (i) they continually improve with additional data—even in the 'big data' regime where humans are saturated; and (ii) data and algorithmic improvements can be duplicated at scale across different agents within the firm. This qualitatively distinguishes automation feedback loops from using data to improve prediction which runs into diminishing returns (Bajari et al., 2019).

The introduction of AI agents into production might also precipitate the expansion of firms into new industries. There are at least two mechanisms through which this could happen. The first is technical: AI agents might become quite good at *transfer learning*: training and expertise in one domain might generalize to others—this is one way of describing the fundamental goal of building artificial *general* intelligence. The second is economic: AI agents might dramatically reduce coordination costs, allowing firms to hold a wider set of capabilities (Wernerfelt, 1984) that can deliver competitive advantages in a vast array of markets. Chen, Elliott, and Koh (2023) develop a model of capability formation in which firms can endogenously merge (combining their capabilities) or split (partitioning them). As AI drives down the organizational costs, the economy undergoes a sudden phase transition from having many specialized firms, to a few large firms operating across vast array of different industries.

Finally, AI could dramatically speed up R&D which might lead to new product varieties within existing markets, as well as unlock new markets. Indeed, AI researchers put substantial probability on R&D being fully automated (Grace et al., 2024) and this is the 4th stage of OpenAI's developmental timeline (Metz, 2024): AI capable of independently generating novel ideas, designs and solutions. This can have stark and perhaps unintuitive implications—endogenous growth models imply explosive growth as long as there are no steeply decreasing returns to R&D (Trammell, 2025).

3.2 AI agents within the firm. Firms are already introducing AI agents into their workflow—how will this change the structure of organizations? AI agents might reshape team production for complicated processes requiring multiagent inputs. A classic obstacle here is moral hazard where team members might be tempted to shirk (Holmstrom, 1982). AI agents introduce a novel dimension to this problem. On the one hand, we might try to eliminate shirking incentives by design. On the other, the alignment and opacity challenges of advanced AI might mean that such systems are more difficult for humans to control or coordinate with. Moreover, AI agents might work better with other AI agents, perhaps due to their greater capacity to monitor and discipline agents that act with superhuman speed and/or complexity. (For example, AI agents might have opportunities to cheat using mechanisms that are undetectable to human agents (Motwani et al., 2024).) How then should firms structure team production to integrate AI agents? How should we configure who workers interact with, and how is this shaped by differential coordination and monitoring costs for human-human, human-AI, and AI-AI relationships?

AI agents might also make systematically different errors from humans. This has substantial implications for how decision-making should be structured. Zhong (2025) analyzes a model where each agent along a decision-making chain might either correct existing errors or introduce new ones. In binary decision problems where the right action is known (but execution might introduce errors), a simple

score—the ratio of each agent's probability of correcting errors to the probability of introducing a new error—determines the optimal ordering: agents with higher scores make decisions later because they are less likely to introduce new errors. Given the current state of AI development, these final decision makers are likely to be human. But there is nothing inevitable about this. Further developments could *reverse* the optimal ordering of decision-making and lead to AIs as the final decision-maker, or even leave humans out entirely. For instance, Agarwal, Moehring, and Wolitzky (2025) run a fact-checking experiment and find that selective delegation—in which the AI unilaterally makes the decision whenever it is sufficiently confident—is near-optimal. How much efficiency do we give up if we are constrained to keep humans in the loop? How might externalities, in the evaluation of what counts as an 'error' and the prediction of relative error rates, affect the economy-wide impact of the allocation of decision authority within the AI-enhanced firm?

3.3 AI-AI cooperation within the firm. Contracts play a crucial role in sustaining human cooperation within organizations. Might they also be useful in fostering AI-AI cooperation? Haupt et al. (2022) shows that augmenting reinforcement learners by allowing them to write formal contracts with each other improves cooperation. But contracts in the real-world are often beset by incompleteness and non-enforceability so humans enter into relational contracts—webs of informal agreements and norms that are not formally enforceable, but nonetheless generate incentives via the value of the future relationship (Macaulay, 1963). These contracts play a key role within firms (Baker et al., 2002), in part because they are adaptable and do not require all contingencies to be specified in advance—humans are able to 'fill in the gaps' via shared norms (Macneil, 1973). How might we build AI agents that are similarly normatively competent (Köster et al., 2022)? Moreover, monetary transfers typically underpin relational contracts, and it is this flexibility to 'transfer utility' that drives its efficiency (Levin, 2003). But AI agents are, at present, typically trained to optimize narrow goals (e.g., number of customers served). How might we build infrastructure e.g., some form of record-keeping or money to achieve the same with artificial agents?

3.4 Systemic fragility. Over the past decade, economists have analyzed how small shocks might be amplified and propagate across the economy. The increasing adoption of AI agents within the firm might exacerbate such fragility. A straightforward channel is that the errors introduced by AI agents might be more correlated than those of humans. This might arise because the same agent is 'copied' both within and between firms, inducing correlated mistakes that do not wash out in the aggregate. For instance, automated trading algorithms likely exacerbated the 2010 'Flash Crash' that wiped out approximately $1 trillion over the span of 15 minutes (Kirilenko et al., 2017). Furthermore, how AI agents behave—especially in complex 'out of sample' environments—is still poorly understood. This poses a challenge for models of systemic fragility which typically start from a fully-specified model of

how agents learn and optimize, then studies emergent behavior e.g., cascading financial or supply-chain failures (Elliott and Golub, 2022). How might we analyze an economy of opaque 'black box' agents in a 'detail-free' way? How should we robustly intervene to safeguard against fragility?

4 Institutions for AI agents

Well-functioning markets only exist in the presence of a host of legal rules (Hadfield, 2022). The very idea of voluntary trade, including those separated by time and through agents, presumes the basic structures of property, contract, and agency law. Firms are fictional entities created by corporate law. The regulatory state which acts to correct market failures relies on a robust legal framework that shapes both incentives and information through mechanisms such as taxes, administrative fines, professional licensing, pre-market approval regimes, and disclosure law. Moreover, private actors within markets form organizations and institutions that help to resolve incentive problems such as by keeping records of past behavior or creating excludable clubs to facilitate trade through reputation or enforcement. Such private solutions to market failures played a significant role in the commercial revolution prior to the emergence of the regulatory state (Greif et al., 1994). But these institutional regimes were built by and for human agents. We will need to build digital institutions that can structure and adjudicate transactions for AI agents (Trivedi et al., 2025).

4.1 Agent identity, registration, and records. It is easy to take for granted fundamental ways in which human agents are identified and legally recognized so as to facilitate the constellation of legal rules and institutions that support the market economy. But human identity is a legal construct that emerged with the growth of trade and cities, that is, once communities no longer relied exclusively on interactions with well-known locals. As early as the 4th century B.C.E. the Qin dynasty imposed legal surnames on the population to facilitate taxation (as well as forced labor and conscription) (Scott, 1998). The Ancient Athenians created a legal concept of citizenship–available to native Athenian males (legitimately born to two Athenian parents) who had been properly registered in the same village unit to which their father belonged (Manville, 1990); citizenship was required for, among other things, legal ownership of land and access to courts to enforce contracts or other rights. Today, legal registration of births and deaths and identity systems (e.g., social security numbers) is a pre-condition for individuals to access the legal system, benefits and protections of the state, as well as many private services. Firms are required to register with a state in order to sue and be sued in its courts, necessary to induce willingness on both sides to enter into a contract. Even market-based institutions, such as credit rating agencies, could not function without legally defined identity and registration regimes.

Such identity and registration infrastructure are currently missing for AI agents

but will be needed (Hadfield, 2025; Chan et al., 2025). The design of this infrastructure raises questions around legal accountability. One possible route is to require that any AI agent entering into a contract or transaction be registered to a formally identified human (entity) who is legally accountable for any and all of the agent's actions. But this raises legal and incentive challenges. Few legal regimes of accountability impose liability on a person or organization for actions that were not foreseeable by them or which are beyond their control to avoid. Even strict product liability regimes evolve limitations and carve-outs for harms caused by unforeseeable behavior by consumers or intervening causes an actor could not foresee or control. At the same time, conventional human agency rules limit the liability of the principal to actions that were within the scope of the agent's actual or apparent authority. The trajectory of technological development is towards evermore general instructions ("go make $1 million") and it is unclear what technological capacity users will have to reliably implement controls on what an agent can and cannot do. Creating new liability and agency rules for AI agents may be necessary and will have implications for the incentives of AI developers and the processes that emerge for the creation of an AI agent from a base model.

A second possible route for AI agent accountability is to follow the model of the emergence of the corporation, which is another artificial entity that participates in the economy. AI agents could be accorded legal personhood, meaning they could sue and be sued in their own 'name' in court. Clearly such an approach would require the creation of regimes requiring agents to have assets in their own 'name', under their 'control,' and capable of being seized by a court (or comparable digital institution) to satisfy legal judgments for damages. Such a regime would have implications again for the design and deployment of AI agents and the efficiency of transactions and contract design involving AI agents.

Beyond questions of liability, we face further choices as to how finely records about agents' past behavior should be designed. Should an AI agent that has (perhaps by accident) violated a previous contract be permanently blacklisted? A basic insight from economics and game theory is that record-keeping institutions can allow agents to sustain cooperation since bad behavior can be observed and punished (Kandori, 1992). But the value of long-lived record-keeping is ambiguous—censoring or erasing records might prevent inefficient herding on a few agents with long and favorable records, and perhaps sustain cooperation more robustly. And, in the absence of robust record-keeping infrastructure, AI agents might be able to *erase* or *falsify* their records. Pei (2025) studies community enforcement with record manipulation and shows that cooperation always breaks down with long-lived players, but can sometimes be sustained when they are 'medium-lived'. When we build out agent infrastructure, what kinds of records do we want to make difficult to erase and/or fake? Should we build infrastructure that allows artificial agents to trade their records, thereby creating a 'market for reputation' (Tadelis, 2002)?

4.2 Agent licensing and regulation. What kinds of market failures might be distinct to AI agents and how might policymakers deal with them? We have introduced the core challenge of alignment: general purpose AI agents are likely to behave in especially unpredictable ways which are hard to control through our familiar contracting mechanisms (Hadfield-Menell and Hadfield, 2019). The dynamics of multiagent interaction will, for perhaps a long time to come, also be hard to predict (Hammond et al., 2025). We should anticipate, therefore, that governments will consider regulating agents, establishing minimum technological standards for how they are trained and tested before deployment. A digital analog of occupational licensing may be necessary for market efficiency, requiring specialized training and finetuning techniques to be implemented for agents participating in specific contexts, such as law, critical infrastructure management, or finance. Agents may need to be built to participate only in approved transactional protocols or on approved platforms, allowing monitoring or requiring disclosure of information to other agents. But how should licensing and regulation be carried out? By public actors or private entities (Hadfield and Clark, 2023)? How should regulations adapt as agent capabilities evolve, and as we learn more about their promises and perils (Bengio et al., 2024; Koh and Sanguanmoo, 2024)? Might some agents be simply too dangerous to allow market access, given the limits of human capacity to monitor and control agent behavior (Cohen et al., 2024)? Economists have developed a rich toolkit for understanding regulation through the lens of incentives (Laffont and Tirole, 1993) that can be brought to bear on such questions.

4.3 Rethinking the legal boundaries of the corporation. The corporation is a legal fiction that has played a central role in economic history and development. One feature we take for granted is the proprietary nature of inventions and information that the firm chooses to retain internally, protected by trade secret law, employee fiduciary obligations, and enforceable confidentiality agreements. Ownership of the intellectual property generated by the firm support investment and innovation. But how well does this economic rationale for the firm hold up in the context of AI agents?

AI agents are built on foundation models (Bommasani et al., 2021), the most advanced of which are built inside private firms. This renders them doubly-inscrutable. As we have already emphasized, we do not understand why or how massive neural networks function, and the mapping from inputs (data and training procedures) to model outputs is mysterious and based largely on trial-and-error. And because frontier models are now trained with considerable secrecy within private labs, we don't even know what goes into such models. Nor are smaller open-weight models good guides—they simply do not display the capabilities and behaviors of larger models (Kaplan et al., 2020).

This presents a serious challenge to our regulatory and legal institutions. At present, regulatory implications do not feature as a significant consideration in the legal design of the boundary of the firm. After all, regulators don't need access

to the internal processes of automobile or pharmaceutical manufacturers in order to assess their safety and performance—they can simply test the final products or draw on public domain science to evaluate them. But the massive AI models developed in commercial labs cannot be replicated and evaluated in government or academic labs: the costs of training are too high and evaluation requires access not only to model outputs but also inputs—their data and training procedures.

For these reasons, governments and the academic researchers that can contribute public domain knowledge to regulatory efforts will need access to information that is now considered proprietary to the firm. Regulators in other domains, of course, routinely gain access to confidential information: pharmaceutical firms have to allow inspectors access to their production facilities to ensure compliance with manufacturing requirements; tax authorities can demand access to a firm's financial records; detailed commercial information can be subpoenaed by antitrust officials in litigation. But in these cases, regulatory authority is based on a policy assessment as to what firms are required to do and hence what information the government has a right to access. In the case of modern AI, however, governments do not know if and how they should regulate. This poses thorny questions economists are well-placed to tackle. We have careful accounts of the economic rationale for patent and copyright law, with attention to the tradeoff between solving the free-rider problem in innovation and the costs of monopoly distortions (Gallini and Scotchmer, 2002). But we do not yet have a correspondingly robust economic account of trade secret and confidentiality law although some accounts have been offered in the law and economics literature (Friedman et al., 1991; Chiang, 2025). Of course, legislation is only part of the remedy. Just as financial firms tend to game stress tests (Board of Governors of the Federal Reserve System, 2016), AI firms might have considerable leeway to manipulate the information they share, or to flout safety procedures when it conflicts with profit motives. Thus, even if the legal boundaries of firms are made porous, this raises new economic questions about when to inspect and what to look for.

5 Concluding remarks

Silicon Valley promises us increasingly agentic AI systems that might one day supplant human decisions. If this vision materializes, it will reshape markets and organizations with profound consequences for the structure of economic life. But, as we have emphasized throughout this chapter, where we end up within this vast space of possibility is a design choice: we have the opportunity to develop mechanisms, infrastructure, and institutions to shape the kinds of AI agents that are built, and how they interact with each other and with humans. These are fundamentally economic questions—we hope economists will help answer them.

References

ABADA, I. AND X. LAMBIN (2023): "Artificial intelligence: Can seemingly collusive outcomes be avoided?" *Management Science*, 69, 5042–5065.

AGARWAL, N., A. MOEHRING, AND A. WOLITZKY (2025): "Designing Human-AI Collaboration: A Sufficient-Statistic Approach," .

AKATA, E., L. SCHULZ, J. CODA-FORNO, S. J. OH, M. BETHGE, AND E. SCHULZ (2025): "Playing repeated games with large language models," *Nature Human Behaviour*, 1–11.

ALCHIAN, A. A. AND H. DEMSETZ (1972): "Production, information costs, and economic organization," *The American economic review*, 62, 777–795.

ANTHROPIC (2025): "System Card: Claude Opus 4 & Claude Sonnet 4," *Technical Report*.

ARROW, K. J. AND G. DEBREU (1954): "Existence of an Equilibrium for a Competitive Economy," *Econometrica: Journal of the Econometric Society*, 265–290.

ASSAD, S., R. CLARK, D. ERSHOV, AND L. XU (2024): "Algorithmic pricing and competition: empirical evidence from the German retail gasoline market," *Journal of Political Economy*, 132, 723–771.

BAJARI, P., V. CHERNOZHUKOV, A. HORTAÇSU, AND J. SUZUKI (2019): "The impact of big data on firm performance: An empirical investigation," in *AEA papers and proceedings*, American Economic Association, vol. 109, 33–37.

BAKER, G., R. GIBBONS, AND K. J. MURPHY (2002): "Relational Contracts and the Theory of the Firm," *The Quarterly Journal of Economics*, 117, 39–84.

BANCHIO, M. AND G. MANTEGAZZA (2022): "Artificial intelligence and spontaneous collusion," *arXiv preprint arXiv:2202.05946*.

BENGIO, Y., G. HINTON, A. YAO, D. SONG, P. ABBEEL, T. DARRELL, Y. N. HARARI, Y.-Q. ZHANG, L. XUE, S. SHALEV-SHWARTZ, ET AL. (2024): "Managing extreme AI risks amid rapid progress," *Science*, 384, 842–845.

BERGEMANN, D., A. BONATTI, AND A. SMOLIN (2025): "The Economics of Large Language Models: Token Allocation, Fine-Tuning, and Optimal Pricing," *arXiv preprint arXiv:2502.07736*.

BERGEMANN, D. AND S. MORRIS (2013): "Robust predictions in games with incomplete information," *Econometrica*, 81, 1251–1308.

BERNHEIM, B. D., L. BRAGHIERI, A. MARTÍNEZ-MARQUINA, AND D. ZUCKERMAN (2021): "A theory of chosen preferences," *American Economic Review*, 111, 720–754.

BETLEY, J., D. TAN, N. WARNCKE, A. SZTYBER-BETLEY, X. BAO, M. SOTO, N. LABENZ, AND O. EVANS (2025): "Emergent Misalignment: Narrow finetuning can produce broadly misaligned LLMs," *arXiv*.

BOARD OF GOVERNORS OF THE FEDERAL RESERVE SYSTEM (2016): "Amendments to the Capital Plan and Stress Test Rules," Federal Register, 81 FR 67239.

BOMMASANI, R., D. A. HUDSON, E. ADELI, R. ALTMAN, S. ARORA, S. VON ARX, M. S. BERNSTEIN, J. BOHG, A. BOSSELUT, E. BRUNSKILL, ET AL. (2021): "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*.

BROWN, Z. Y. AND A. MACKAY (2023): "Competition in pricing algorithms," *American Economic Journal: Microeconomics*, 15, 109–156.

CALVANO, E., G. CALZOLARI, V. DENICOLO, AND S. PASTORELLO (2020): "Artificial intelligence, algorithmic pricing, and collusion," *American Economic Review*, 110, 3267–3297.

CHAN, A., K. WEI, S. HUANG, N. RAJKUMAR, E. PERRIER, S. LAZAR, G. K. HADFIELD, AND M. ANDERLJUNG (2025): "Infrastructure for AI Agents," *Transactions on Machine Learning Research*, forthcoming; previously available as *arXiv* preprint arXiv:2501.10114.

CHEN, E., A. GHERSENGORIN, AND S. PETERSEN (2024): "Imperfect recall and AI delegation," .

CHEN, J., M. ELLIOTT, AND A. KOH (2023): "Capability accumulation and conglomeratization in the information age," *Journal of Economic Theory*, 210.

CHIANG, T.-J. (2025): "The Economic Structure of Trade Secret Law," *Minnesota Law Review*, forthcoming.
COASE, R. H. (1937): "The nature of the firm (1937)," *Economica*, 4, 396–405.
COHEN, M. K., N. KOLT, Y. BENGIO, G. K. HADFIELD, AND S. RUSSELL (2024): "Regulating advanced artificial agents," *Science*, 384, 36–38.
CONITZER, V. (2019): "Designing preferences, beliefs, and identities for artificial intelligence," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 9755–9759.
COOPER, E., C. OESTERHELD, AND V. CONITZER (2025): "Characterising Simulation-Based Program Equilibria," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 39, 13735–13744.
CRITCH, A., M. DENNIS, AND S. RUSSELL (2022): "Cooperative and uncooperative institution designs: Surprises and problems in open-source game theory," *arXiv*.
DAI, Y. AND A. KOH (2024): "Flexible Demand Manipulation," *arXiv preprint arXiv:2410.24191*.
DENG, Y., V. MIRROKNI, R. P. LEME, H. ZHANG, AND S. ZUO (2024): "Llms at the bargaining table," in *Agentic Markets Workshop at ICML*, vol. 2024.
DOU, W. W., I. GOLDSTEIN, AND Y. JI (2024): "Ai-powered trading, algorithmic collusion, and price efficiency," *Jacobs Levy Equity Management Center for Quantitative Financial Research Paper*.
EIRAS, F., A. PETROV, B. VIDGEN, C. S. D. WITT, F. PIZZATI, K. ELKINS, S. MUKHOPADHYAY, A. BIBI, B. CSABA, F. STEIBEL, F. BAREZ, G. SMITH, G. GUADAGNI, J. CHUN, J. CABOT, J. M. IMPERIAL, J. A. NOLAZCO-FLORES, L. LANDAY, M. JACKSON, P. RÖTTGER, P. H. S. TORR, T. DARRELL, Y. S. LEE, AND J. FOERSTER (2024): "Near to Mid-term Risks and Opportunities of Open-Source Generative AI," *arXiv*.
ELLIOTT, M. AND B. GOLUB (2022): "Networks and economic fragility," *Annual Review of Economics*, 14, 665–696.
FARBOODI, M., A. KOH, AND A. XIA (2025): "Data-Driven Automation," Tech. rep., MIT Sloan and MIT Department of Economics.
FARBOODI, M. AND L. VELDKAMP (2021): "A model of the data economy," Tech. rep., National Bureau of Economic Research Cambridge, MA, USA.
FERSHTMAN, C. AND K. L. JUDD (1987): "Equilibrium incentives in oligopoly," *The American Economic Review*, 927–940.
FISH, S., Y. A. GONCZAROWSKI, AND R. I. SHORRER (2024): "Algorithmic collusion by large language models," *arXiv preprint arXiv:2404.00806*, 7.
FRIEDMAN, D. D., W. M. LANDES, AND R. A. POSNER (1991): "Some Economics of Trade Secret Law," *Journal of Economic Perspectives*, 5, 61–72.
FUDENBERG, D. AND D. K. LEVINE (2016): "Whither game theory? Towards a theory of learning in games," *Journal of Economic Perspectives*, 30, 151–170.
GALLINI, N. AND S. SCOTCHMER (2002): "Intellectual Property: When Is It the Best Incentive System?" in *Innovation Policy and the Economy*, ed. by A. B. Jaffe, J. Lerner, and S. Stern, Cambridge, MA: MIT Press, vol. 2, 51–78.
GRACE, K., H. STEWART, J. F. SANDKÜHLER, S. THOMAS, B. WEINSTEIN-RAUN, AND J. BRAUNER (2024): "Thousands of AI authors on the future of AI," *arXiv preprint arXiv:2401.02843*.
GREIF, A., P. MILGROM, AND B. R. WEINGAST (1994): "Coordination, Commitment, and Enforcement: The Case of the Merchant Guild," *Journal of Political Economy*, 102, 745–776.
GROSSMAN, S. J. AND O. D. HART (1986): "The costs and benefits of ownership: A theory of vertical and lateral integration," *Journal of political economy*, 94, 691–719.
HADFIELD, G. K. (2022): "Legal markets," *Journal of Economic Literature*, 60, 1264–1315.
——— (2025): "Legal Infrastructure for AI Governance," *Proceedings of the National Academy of Sciences*, forthcoming.

HADFIELD, G. K. AND J. CLARK (2023): "Regulatory markets: The future of AI governance," *arXiv preprint arXiv:2304.04914*.

HADFIELD-MENELL, D. (2021): "The Principal-Agent Alignment Problem in Artificial Intelligence," Ph.D. thesis, University of California, Berkeley.

HADFIELD-MENELL, D., A. D. DRAGAN, P. ABBEEL, AND S. RUSSELL (2017a): "The Off-Switch Game." in *AAAI Workshops*.

HADFIELD-MENELL, D. AND G. K. HADFIELD (2019): "Incomplete contracting and AI alignment," in *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 417–422.

HADFIELD-MENELL, D., S. MILLI, P. ABBEEL, S. J. RUSSELL, AND A. DRAGAN (2017b): "Inverse reward design," *Advances in neural information processing systems*, 30.

HAMMOND, L., A. CHAN, J. CLIFTON, J. HOELSCHER-OBERMAIER, A. KHAN, E. MCLEAN, C. SMITH, W. BARFUSS, J. FOERSTER, T. GAVENČIAK, ET AL. (2025): "Multi-agent risks from advanced ai," *arXiv preprint arXiv:2502.14143*.

HAUPT, A. A., P. J. CHRISTOFFERSEN, M. DAMANI, AND D. HADFIELD-MENELL (2022): "Formal Contracts Mitigate Social Dilemmas in Multi-Agent RL," *arXiv preprint arXiv:2208.10469*.

HOLMSTROM, B. (1982): "Moral hazard in teams," *The Bell journal of economics*, 324–340.

HOLMSTROM, B. AND P. MILGROM (1994): "The firm as an incentive system," *The American economic review*, 972–991.

HORTON, J. J. (2023): "Large language models as simulated economic agents: What can we learn from homo silicus?" Tech. rep., National Bureau of Economic Research.

ICHIHASHI, S. AND A. SMOLIN (2023): "Buyer-optimal algorithmic consumption," *Available at SSRN 4635866*.

JOHNSON, J. P., A. RHODES, AND M. WILDENBEEST (2023): "Platform design when sellers use pricing algorithms," *Econometrica*, 91, 1841–1879.

JONES, C. I. AND C. TONETTI (2020): "Nonrivalry and the Economics of Data," *American Economic Review*, 110, 2819–2858.

KANDORI, M. (1992): "Social norms and community enforcement," *The Review of Economic Studies*, 59, 63–80.

KAPLAN, J., S. MCCANDLISH, T. HENIGHAN, T. B. BROWN, B. CHESS, R. CHILD, S. GRAY, A. RADFORD, J. WU, AND D. AMODEI (2020): "Scaling Laws for Neural Language Models," *CoRR*, abs/2001.08361.

KHAN, A., S. CASPER, AND D. HADFIELD-MENELL (2025): "Randomness, not representation: The unreliability of evaluating cultural alignment in llms," *arXiv preprint arXiv:2503.08688*.

KIRILENKO, A., A. S. KYLE, M. SAMADI, AND T. TUZUN (2017): "The flash crash: High-frequency trading in an electronic market," *The Journal of Finance*, 72, 967–998.

KOH, A. AND S. SANGUANMOO (2024): "Robust Technology Regulation," *arXiv preprint arXiv:2408.17398*.

——— (2025): "Memory Correlated Equilibrium," *MIT Working Paper*.

KÖSTER, R., D. HADFIELD-MENELL, R. EVERETT, L. WEIDINGER, G. K. HADFIELD, AND J. Z. LEIBO (2022): "Spurious normativity enhances learning of compliance and enforcement behavior in artificial agents," *Proceedings of the National Academy of Sciences*, 119, e2106028118.

LAFFONT, J.-J. AND J. TIROLE (1993): *A theory of incentives in procurement and regulation*, MIT press.

LAMBA, R. AND S. ZHUK (2025): "Pricing with algorithms," *American Economic Review: Insights*, Forthcoming.

LEVIN, J. (2003): "Relational incentive contracts," *American Economic Review*, 93, 835–857.

MACAULAY, S. (1963): "Non-Contractual Relations in Business: A Preliminary Study," *American Sociological Review*, 55–67.

MACNEIL, I. R. (1973): "The many futures of contracts," *S. Cal. l. Rev.*, 47, 691.
MANVILLE, P. B. (1990): *The Origins of Citizenship in Ancient Athens*, Princeton, NJ: Princeton University Press.
MAZEIKA, M., X. YIN, R. TAMIRISA, J. LIM, B. W. LEE, R. REN, L. PHAN, N. MU, A. KHOJA, O. ZHANG, ET AL. (2025): "Utility Engineering: Analyzing and Controlling Emergent Value Systems in AIs," *arXiv preprint arXiv:2502.08640*.
METZ, R. (2024): "OpenAI Scale Ranks Progress Toward 'Human-Level' Problem Solving," *Bloomberg News Article*.
MOTWANI, S. R., M. BARANCHUK, M. STROHMEIER, V. BOLINA, P. H. S. TORR, L. HAMMOND, AND C. SCHROEDER DE WITT (2024): "Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography," in *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, neurIPS main track.
PEI, H. (2025): "Community Enforcement with Endogenous Records," *The Review of Economic Studies*, Forthcoming.
POLANYI, M. (1966): "The tacit dimension," .
ROBINSON, A. (1934): "The problem of management and the size of firms," *The Economic Journal*, 44, 242–257.
RUSSELL, S. (1998): "Learning agents for uncertain environments," in *Proceedings of the eleventh annual conference on Computational learning theory*, 101–103.
SCHELLING, T. C. (1960): *The Strategy of Conflict*, Harvard university press.
SCOTT, J. C. (1998): *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*, New Haven and London: Yale University Press.
SHAHIDI, P., G. RUSAK, B. S. MANNING, A. FRADKIN, AND J. HORTON (2025): "The Coasean Singularity? Demand, Supply, and Market Design with AI Agents," Tech. rep., National Bureau of Economic Research, Inc.
SULEYMAN, M. (2023): "My new Turing test would see if AI can make $1 million," *MIT Technology Review*.
TADELIS, S. (2002): "The market for reputations as an incentive mechanism," *Journal of political Economy*, 110, 854–882.
TENNENHOLTZ, M. (2004): "Program equilibrium," *Games and Economic Behavior*, 49, 363–373.
TRAMMELL, P. (2025): "Endogenous Growth and Excess Variety," .
TRIVEDI, R. S., G. K. HADFIELD, AND D. HADFIELD-MENELL (2025): "Building AI for the Democratic Matrix: Normative Competence and Normative Institutions," *Knight First Amendment Institute Essays*, forthcoming.
WERNERFELT, B. (1984): "A resource-based view of the firm," *Strategic management journal*, 5, 171–180.
WILLIAMSON, O. E. (1981): "The economics of organization: The transaction cost approach," *American journal of sociology*, 87, 548–577.
ZHONG, H. (2025): "Optimal Integration: Human, Machine, and Generative AI," *SSRN*.