REVIEW

Economic reasoning and artificial intelligence

David C. Parkes[1]* and Michael P. Wellman[2]*

The field of artificial intelligence (AI) strives to build rational agents capable of perceiving the world around them and taking actions to advance specified goals. Put another way, AI researchers aim to construct a synthetic *homo economicus*, the mythical perfectly rational agent of neoclassical economics. We review progress toward creating this new species of machine, *machina economicus*, and discuss some challenges in designing AIs that can reason effectively in economic contexts. Supposing that AI succeeds in this quest, or at least comes close enough that it is useful to think about AIs in rationalistic terms, we ask how to design the rules of interaction in multi-agent systems that come to represent an economy of AIs. Theories of normative design from economics may prove more relevant for artificial agents than human agents, with AIs that better respect idealized assumptions of rationality than people, interacting through novel rules and incentive systems quite distinct from those tailored for people.

Economics models the behavior of people, firms, and other decision-makers as a means to understand how these decisions shape the pattern of activities that produce value and ultimately satisfy (or fail to satisfy) human needs and desires. In this enterprise, the field classically starts from an assumption that actors behave rationally—that is, their decisions are the best possible given their available actions, their preferences, and their beliefs about the outcomes of these actions. Economics is drawn to rational decision models because they directly connect choices and values in a mathematically precise manner. Critics argue that the field studies a mythical species, *homo economicus* ("economic man") and produces theories with limited applicability to how real humans behave. Defenders acknowledge that rationality is an idealization but counter that the abstraction supports powerful analysis, which is often quite predictive of people's behavior (as individuals or in aggregate). Even if not perfectly accurate representations, rational models also allow preferences to be estimated from observed actions and build understanding that can usefully inform policy.

Artificial intelligence (AI) research is likewise drawn to rationality concepts, because they provide an ideal for the computational artifacts it seeks to create. Core to the modern conception of AI is the idea of designing agents: entities that perceive the world and act in it (*1*). The quality of an AI design is judged by how well the agent's actions advance specified goals, conditioned on the perceptions observed. This coherence among perceptions, actions, and goals is the essence of rationality. If we represent goals in terms of preference over outcomes, and conceive perception and action within the framework of decision-making under uncertainty, then the AI agent's situation aligns squarely with the standard economic paradigm of rational choice. Thus, the AI designer's task is to build rational agents, or agents that best approximate rationality given the limits of their computational resources (*2*–*4*). In other words, AI strives to construct—out of silicon (or whatever) and information—a synthetic *homo economicus*, perhaps more accurately termed *machina economicus*.

The shared rationality abstraction provides a strong foundation for research that spans AI and economics. We start this review by describing progress on the question of how to operationalize rationality and how to construct AI agents that are able to reason about other AIs. Supposing that AI research succeeds in developing an agent that can be usefully modeled as rational (perhaps more so than human agents), we turn to research on the design of systems populated by multiple AIs. These multi-agent systems will function as AI economies, with AIs engaged in transactions with other AIs as well as with firms and people. This prospect has spawned interest in expanding theories of normative design from economics, optimizing rules of encounter (*5*) to guide multi-agent interactions. Systems populated by AIs may exhibit new economic phenomena and thus require a new science with which to understand the way they function and to guide their design. For example, although human cognitive constraints limit the design of current markets, systems designed for AIs may admit more complex interfaces, impose greater calculation burdens, and demand more stamina of attention.

At the same time, the ways in which the behavior of AIs deviate from the behavior of people can present new challenges. We can already glimpse the future of economic AIs, with simple AI bots pricing books for sale on Amazon and scanning for restaurant tables on Opentable for resale at a profit (*6*). Such AIs may introduce some efficiencies, but their lack of common sense and their designer's failure to anticipate interactions can also lead to books priced at $23 million (*7*). More sophisticated AI strategies, presumably more carefully vetted, exert a large influence on financial markets, with automated trading algorithms estimated to be responsible for more than 70% of trades on U.S. stock markets (*8*). Given the consequences, it is important to understand the effect of ubiquitous automated agents on the performance of economic systems. As reasoning is shifted from people to AIs—designed to learn our preferences, overcome our decision biases, and make complex cost-benefit trade-offs—how too should the economic institutions that mediate everyday transactions change?

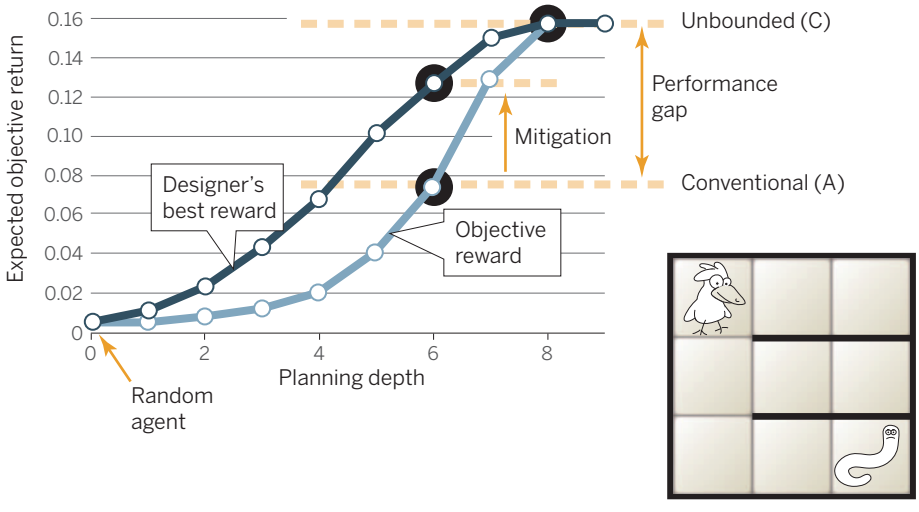

Fig. 1. A bounded reinforcement learning agent performs better by pursuing a designed reward function different from the objective reward: its actual fitness evaluation. Results (left) from a gridworld foraging domain (right), for various limits on the agent's planning horizon (*84*). Unless the agent is perfectly rational (i.e., no horizon limit)—not typically feasible in realistic applications—the designer can often achieve better fitness by directing the agent to optimize an alternative measure.

[1]Harvard John A. Paulson School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge MA 02138, USA. [2]Computer Science and Engineering, University of Michigan, 2260 Hayward Street, Ann Arbor, MI 48109, USA.
*Corresponding author. E-mail: parkes@eecs.harvard.edu (D.C.P.); wellman@umich.edu (M.P.W.)

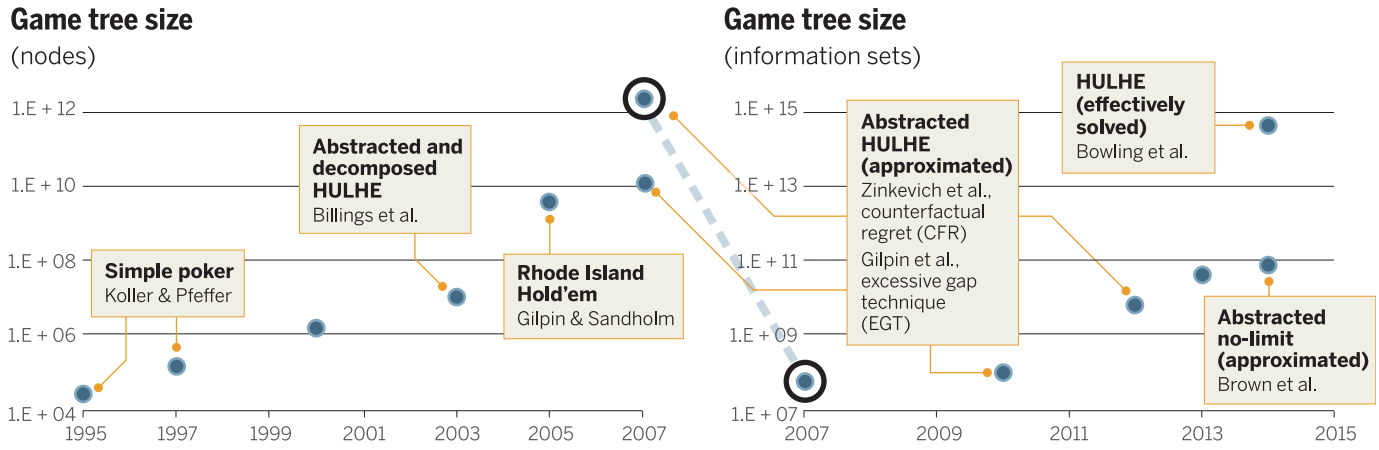

Fig. 2. Researchers produced steady exponential progress on solving games of imperfect information from 1995 to the present. Up to 2007 (left), game size was generally reported in terms of nodes in the game tree. Based on methods introduced around that time, it became more meaningful (right) to report size in terms of the number of information sets (each many nodes), which represent distinct situations as perceived from the perspective of a player. The circled data points correspond to the same milestone; combining the two graphs thus demonstrates the continual exponential improvement. Data are from (*23*, *35*, *85*–*90*).

We focus here on some of the research directions we consider most salient for a future synthesis of economics and AI engendered by the emergence of *machina economicus*. Interesting as they are, we only briefly mention here the many exciting applications of AI to problems in economics such as matching (*9*), market clearing (*10*), and preference modeling for smart grids (*11*). Nor will we showcase the many ways in which economic theory is finding application today within AI—for example, game-theoretic approaches to multi-agent learning (*12*) and voting procedures to combine the opinions of AIs (*13*).

Building *machina economicus*

Constructing a rational AI raises a host of technical challenges not previously addressed in the long tradition of rationalistic modeling in the social sciences. For economics, the agent attitudes (e.g., beliefs and preferences) underlying rationality are conceptual abstractions. Economists need not explain how capabilities and preferences, for example, are encoded, nor the algorithm by which an agent plans what actions to take conditional on its perceptions. Computation is abstracted away in the standard economic model and is precisely what the AI scientist must account for to operationalize rationality in a realized agent.

This does not mean that an AI design needs to incorporate data structures corresponding directly to rationality constructs, although many AI architectures do feature direct representations for propositions, goals, and the like. Such representations may simplify the analysis of AI systems—for example, we can ask whether an inference algorithm operating on logical expressions possesses desirable properties such as soundness: that all conclusions follow from the premises. Similarly, if an AI's beliefs are encoded as probability distributions, we can ask whether it updates its beliefs from observations in proper accord with Bayesian theory. However, care must be taken in understanding an agent's attitudes solely in terms of its internal data structures. Imperfections in decision-making may mean that the beliefs held and objectives pursued by a computational agent, in effect, vary systematically from those directly encoded.

As an example illustrating this distinction, machine-learning researchers adapted from animal learning the concept of reward shaping (*14*). In reinforcement learning, the agent derives a policy (mapping from perception sequences to actions) based on rewards representing instantaneous value associated with a state and action. A designer specifying the input reward can often train the agent more efficiently by shaping the reward signal over the learning process to facilitate convergence to behavior optimizing the designer's objective. The framework of optimal rewards (*15*) provides a general treatment distinguishing reward specifications and designer goals. As shown in Fig. 1, the optimal reward input to the agent does not generally correspond to the designer's ideal reward. This perspective helps explain the role of intrinsic motivations (e.g., curiosity) in a flexible learning agent.

Although the mantle of designing *machina economicus* may not be adopted (particularly in such explicit terms) by all AI researchers, many AI advances over the past few decades can be characterized as progress in operationalizing rationality. For instance, probabilistic reasoning was largely eschewed by AI 30 years ago but now pervades the field, thanks to developments in representation and inference using Bayesian networks and related graphical formalisms. Expressing uncertainty about general relationships, beyond mere propositions, is routinely supported in probabilistic modeling languages (*16*). Statistical approaches now dominate machine learning and natural language processing (*17*, *18*). Likewise, preference handling (including methods for eliciting preferences from the designer of an AI agent, compactly representing preferences over complex domains, and enabling inference about preferences) is regarded as a necessary AI facility. Planning, the AI subfield concerned with action over time, now conventionally frames its problem as one of optimization, subject to resource constraints, multiple objectives, and probabilistic effects of actions.

Will AI succeed in developing the ideal rational agent? As much as we strive to create *machina economicus*, absolutely perfect rationality is unachievable with finite computational resources. A more salient question is whether AI agents will be sufficiently close to the ideal as to merit thinking about them and interacting with them in rationalistic terms. Such is already the case, at least in a limited sense. Whenever we anthropomorphize our machines, we are essentially treating them as rational beings, responding to them in terms of our models of their knowledge, goals, and intentions. A more refined version of the question is whether our formal rationality theories will fit well the behavior of AI agents in absolute terms or compared to how well the theories work for people. Without offering any judgment on the question of how well rationality theories capture essential human behavior, we note the irony in the prospect that social science theories may turn out to apply with greater fidelity to nonhuman agent behavior.

Reasoning about other agents

The issue of agent theorizing is not merely academic. If we can build one AI agent, then we can build many, and these AIs will need to reason about each other as well as about people. For AIs designed to approximate *machina economicus*, it stands to reason that they should treat each other as rational, at least as a baseline assumption. These AIs would adopt a game-theoretic view of the world, where agents rationally respond to each others' behavior, presumed (recursively) to be rational as well. A consequence is that agents would expect their joint decisions to be in some form of equilibrium, as in standard economic thinking.

That AIs (or AI-human combinations) are reasonably modeled as approximately rational is the premise of a growing body of AI research applying economic equilibrium models to scenarios involving multiple agents (*19*). The approach has achieved notable successes, providing evidence for the premise, at least in particular circumstances. Just as single-agent rationality does not require literal expected-utility calculations, applicability of an equilibrium model does not require that agents themselves be explicitly engaged in equilibrium reasoning. For example, the literature on learning in games (*20*) has identified numerous conditions in which simple adaptive strategies converge to strategic equilibria. We can evaluate the effectiveness of economic modeling by examining agents built by AI designers for specified tasks. For instance, in a study of AI trading agents competing in a shopping game (*21*), an agent using standard price equilibrium models from economics (specifically, Walrasian equilibrium) achieved comparable prediction accuracy to sophisticated machine-learning approaches without using any data, even though none of the other agents employed equilibrium reasoning.

A Prisoner's dilemma

	Cooperate	Defect
Cooperate	4,4	0,6
Defect	6,0	1,1

B Mediated Prisoner's dilemma

	Mediator	Cooperate	Defect
Mediator	4,4	6,0	1,1
Cooperate	0,6	4,4	0,6
Defect	1,1	6,0	1,1

Fig. 3. Each entry gives the utility to (row player, column player). (**A**) Prisoner's dilemma. The dominant strategy equilibrium is (Defect, Defect). (**B**) Mediated prisoner's dilemma. The dominant strategy equilibrium is (Mediator, Mediator).

In the rest of this section, we describe further examples in which economic modeling, in the form of game-theoretic algorithms, has provided an effective way for AIs to reason about other agents. The first example is computer poker. Although poker is an artificial game, many humans have invested a great deal of time and money to develop their playing skills. More important, poker's uncertainty and complexity have made it a compelling challenge problem for AI techniques. Early approaches aimed to capture the knowledge of expert human players (*22*), but over the past decade, game-theoretic algorithms have predominated. Technically, poker is a game of imperfect information, where each player knows elements of history (cards dealt to them) that are secret from others. As uncertainty gets partially resolved over time, through card turns and betting, players must update their beliefs about both card outcomes and the beliefs of others.

A major milestone in computer poker was achieved in 2014 with the effective solution of "heads up limit hold'em" (HULHE), which is a standard two-player version of the most popular poker game (*23*). HULHE is the largest game of imperfect information ever solved (with more than 10^{13} information sets after removing symmetries) and the first imperfect-information game widely played by humans to be solved. The solution was the culmination of two decades of effort by a series of researchers (see Fig. 2), beginning with the exact solution of simplified poker games, and proceeding to the approximate solution of abstracted versions of the full game (*24*). Computing the approximate Nash equilibrium of the full game required massive computation and new methods for equilibrium search based on regret-matching techniques from machine learning. The result is a strategy against which even a perfect opponent cannot earn a detectable profit.

In general, the optimal strategy against perfect opponents may not be the ideal strategy against the more typical fallible kind. Despite considerable effort, however, researchers have not found poker algorithms that perform considerably better than game-theoretic solutions, even against natural distributions of opponents. It has also turned out that game-theoretic approaches have been more successful than alternatives, even for poker variants that are far from being exactly solved, such as no-limit (where bets are unrestricted) (*25*), or games with three or more players (*26*).

Much of the interest in game-theoretic reasoning for AI is driven by its applicability to real-world problems. The most prominent area of application in recent years, and our second example, is that of security games, based on a pioneering series of systems developed by Tambe *et al.* (*27*). In these systems, an agent decides how to defend facilities (e.g., airport security through placement of checkpoints) by solving a game where an attacker is presumed to rationally plan in response to the defender's decision. This approach has been successfully deployed in a variety of domains, including airport and airline security and coast guard patrols.

As for any game-theoretic approach, the recommendations from these systems are sensitive to assumptions made about the other agents (here, attackers): their respective preferences, beliefs, capabilities, and level of rationality. Representational approaches from AI provide flexibility, allowing the assumptions made in the strict versions typically employed by game theorists to be relaxed (*28*). The field of behavioral game theory has developed detailed predictive models based on how humans have been observed to deviate from game-theoretic rationality (*29*). Such predictive models can be readily incorporated in existing game-theoretic reasoning algorithms, as has been demonstrated in the context of modeling attackers in security games (*30*). An interesting open question is whether the kinds of behavioral models that best explain human decision-making [see Wright and Leyton-Brown (*31*) for a meta-study] will also prove effective in capturing the bounded rationality of computational agents.

Designing multi-agent systems

At the multi-agent level, a designer cannot directly program behavior of the AIs but instead defines the rules and incentives that govern interactions among AIs. The idea is to change the "rules of the game" (e.g., rewards associated with actions and outcomes) to effect change in agent behavior and achieve system-wide goals. System goals might include, for instance, promoting an allocation of resources to maximize total value, coordinating behavior to complete a project on time, or pooling decentralized information to form an accurate prediction about a future event. The power to change the interaction environment is special and distinguishes this level of design from the standard AI design problem of performing well in the world as given.

An interesting middle ground is to take the world as given but employ reliable entities—mediators—that can interact with AIs and perform actions on their behalf (*32*). Introducing mediating entities is relatively straightforward in the new AI economy. To see how this can be powerful, consider a mediated extension of the classic prisoner's dilemma game (Fig. 3). If both AIs grant the mediator the authority to play on their behalf (i.e., proxy right), it performs Cooperate on behalf of both agents. However, if only one AI grants the mediator proxy, it performs Defect on behalf of that agent. In equilibrium, both AIs grant proxy, and the effect is to change the outcome from (Defect, Defect) to (Cooperate, Cooperate), increasing utility to both participants.

For the more general specification of rules of interaction for rational agents, economics has a well-developed mathematical theory of mechanism design (*33*). The framework of mechanism design has been fruitfully applied, for example, to the design of matching markets (*34*) and auctions (*35*). Mechanism design is a kind of inverse game theory, with the rules inducing a game and the quality of the system evaluated in an equilibrium. In the standard model, design goals are specified in terms of agent preferences on outcomes, but these preferences are private and the agents are self-interested. A mechanism is a trusted entity, able to receive messages from agents that make claims (perhaps untruthfully) about preferences and select an outcome (e.g., an allocation of resources or a plan of behavior) on the basis of these messages. The challenge is to align incentives and promote truthful reports.

Varian (*36*) has argued that the theory of mechanism design may actually prove more relevant for artificial agents than for human agents, because AIs may better respect the idealized assumptions of rationality made in this framework. For example, one desirable property of a mechanism is incentive compatibility, which stipulates that truthful reports constitute an equilibrium. Sometimes it is even possible to make truthful reporting a dominant strategy (optimal whatever others do), achieving the strong property of strategy-proofness (*37*). It seems, however, that people do not reliably understand this property; evidence from medical matching markets, and also from laboratory experiments, suggests that some participants in strategy-proof matching mechanisms try to misrepresent their preferences even though it provides no advantage (*38*, *39*).

For artificial systems, in comparison, we might expect AIs to be truthful where this is optimal and to avoid spending computation reasoning about the behavior of others where this is not useful (*5*). More generally, mechanism designs for AI systems need not be simple because they need not be understandable to people. On the contrary, AI techniques such as preference representation, preference elicitation, and search algorithms can be used to turn the mathematical formalisms of mechanism design into concrete computational methods (*40*–*42*). The design problem itself can also be usefully formulated as a computational problem, with optimization and machine learning used to find solutions to design problems for which analytical solutions are unavailable (*43*–*46*).

The prospect of an economy of AIs has also inspired expansions to new mechanism design settings. Researchers have developed incentive-compatible multiperiod mechanisms, considering such factors as uncertainty about the future and changes to agent preferences because of changes in local context (*47*–*49*). Another direction considers new kinds of private inputs beyond preference information (*50*, *51*). For example, in a team formation setting, each AI might misreport information about the capabilities of other AIs in order to get itself selected for the team (*52*). Similarly, AIs seeking to maximize task assignments might provide false reports of experience in task performance in order to mislead a learning mechanism constructing an automatic task classifier (*53*). Systems of AIs can also create new challenges for mechanism design. One such challenge is false-name bidding, where an AI exploits its ability to manage multiple identities. For example, it may gain resources more cheaply by dividing a request into a set of smaller requests, each placed from a different identity under its control. In response, researchers have developed mechanisms that are robust to this new kind of attack (*54*).

The important role of mechanism design in an economy of AIs can be observed in practice. Search engines run auctions to allocate ads to positions alongside search queries. Advertisers bid for their ads to appear in response to specific queries (e.g., "personal injury lawyer"). Ads are ranked according to bid amount (as well as other factors, such as ad quality), with higher-ranked ads receiving a higher position on the search results page. Early auction mechanisms employed first-price rules, charging an advertiser its bid amount when its ad receives a click. Recognizing this, advertisers employed AIs to monitor queries of interest, ordered to bid as little as possible to hold onto the current position. This practice led to cascades of responses in the form of bidding wars, amounting to a waste of computation and market inefficiency (*55*). To combat this, search engines introduced second-price auction mechanisms (*37*), which charge advertisers based on the next-highest bid price rather than their own price. This approach (a standard idea of mechanism design) removed the need to continually monitor the bidding to get the best price for position, thereby ending bidding wars (*56*).

In recent years, search engine auctions have supported richer, goal-based bidding languages. For example, an advertiser can ask to maximize clicks over a weighted set of queries subject to a budget constraint (*57*, *58*). Search engines provide proxy agents that then bid on behalf of advertisers to achieve the stated goal (*59*). This introduction of proxy agents and the earlier switch from first price to second price can be interpreted as a computational application of a fundamental concept in mechanism design—the revelation principle (*60*–*62*). Briefly, this states that if the rules of a mechanism and the equilibrium strategies in that mechanism are replaced by a new mechanism that is functionally equivalent to the composition of these rules and strategies, then the new mechanism will be incentive compatible. Although neither redesign provides incentive compatibility in a formal sense, both second-pricing and proxy bidding can be interpreted as accomplishing on behalf of advertisers what they were doing (through AIs) in an earlier design (see Fig. 4). Still other ad platform designs are using a strategy-proof mechanism [the Vickrey-Clarke-Groves mechanism (*37*, *63*, *64*)] to make decisions about the space to allocate to ads, which ads to allocate, and which (nonsponsored) content to display to a user (*65*).

The tangle between automated agents and the design of rules of interaction also features prominently in today's financial markets, where the dominance of computerized traders has, by most accounts, qualitatively shaped the behavior of these markets. Although details of implementation are closely held secrets, it is well understood that techniques from AI and machine learning are widely employed in the design and analysis of algorithmic traders (*66*). Algorithmic trading has enabled the deployment of strategies that exploit speed advantages and has led in turn to a costly arms race of measures to respond to market information with minimum latency. A proposed design response would replace continuous-time auctions with periodic auctions that clear on the order of once per second, thus negating the advantage of tiny speed improvements (*67*, *68*).

We describe two additional examples of the design of multi-agent systems for an economy of AIs. The first example system aggregates information held by multiple AIs. The rules of a system that achieves this goal can be engineered purposefully through the design of a prediction market (*69*). Popular versions of prediction markets feature questions such as who will be elected U.S. president (e.g., Betfair offers many such markets). The basic idea of a prediction market is to facilitate trade in securities contracts (e.g., a possible contract will pay $1 if Hilary Clinton is elected). The price that balances supply and demand is then interpreted as a market prediction (e.g., price $0.60 reflects probability 0.6 for the payoff event).

Consider a domain with a large number of interrelated random variables—for example, "flight BA214 delayed by more than 1 hour," "snowstorm in Boston," "de-icing machine fail," "incoming flight BA215 delayed by more than 1 hour," and "security alert in London." In a combinatorial prediction market (*70*), a large bet on the contract "de-icing machine fail" would affect the price of "flight BA214 delayed by more than 1 hour" and all other connected events. A challenge is that the number of conceivable events is exponential in the number of random variables. Among other properties, a good market design should allow bets

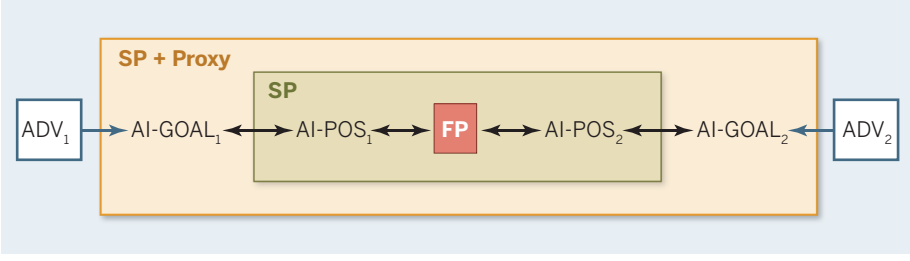

Fig. 4. Two generations of sponsored search mechanisms. Early designs were first price (FP), and advertisers (ADV) used AIs (AI-POS) to maintain a position on the list of search results at the lowest possible price. Second-price (SP) auction mechanisms were introduced, designed to replace the combination of FP and AI-POS. Advertisers adopted new AIs (AI-GOAL) to achieve higher-level goals such as to maximize profit or to maximize the number of clicks. The second price auction was extended to include proxy agents (SP+Proxy), designed to replace the combination of SP and AI-GOAL.

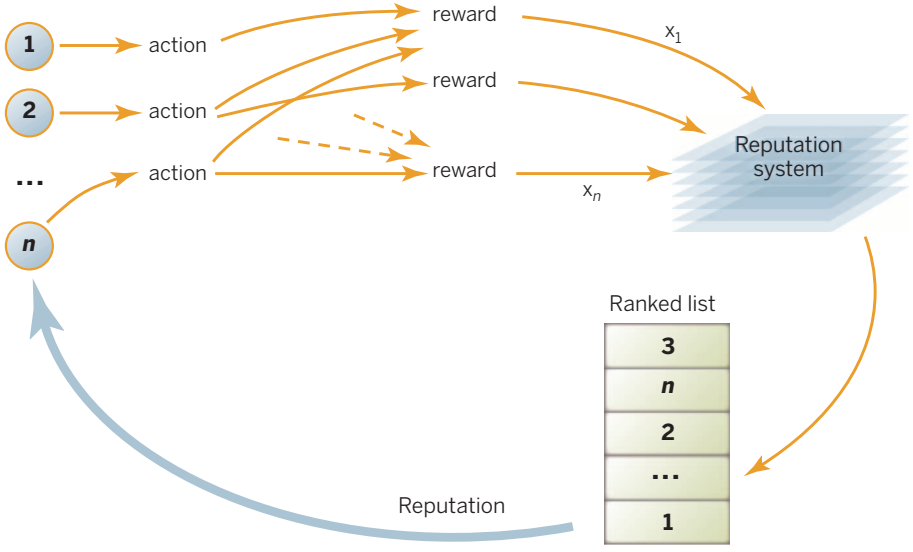

Fig. 5. In a reputation system for a multi-agent AI, each agent chooses an action, and the combined effect of these actions generates rewards (i.e., utility). Based on the actions taken and the rewards received, agent i can submit a report, xi, to the reputation system. The reputation system aggregates this feedback—for example, providing a ranked list to reflect the estimated trustworthiness of agents. Each agent observes this ranked list, and this information may influence future actions.

on all events about which AIs have information (e.g., "de-icing machine fail AND all subsequent flights from Boston delayed by more than 1 hour"). A good design should also align incentives—for example, making it utility-maximizing to trade immediately on current information until the market price reflects an agent's belief. Progress in scaling up combinatorial markets has been made by relating the problem of pricing bets to well-understood problems in statistical inference and convex optimization (*71*, *72*). Related research advances are being made by allowing AIs to transact in hypotheses that are acquired through machine learning as well as trade directly in information signals rather than beliefs (*73*–*75*).

The second example is the management of information concerning the trustworthiness of agents within an economy of AIs. Trust that a counterparty will complete a transaction or invest effort or resources is crucial for any well-functioning economic system. A standard approach is to associate participants with a reputation, which can serve to align incentives in the present under the threat of a damaged reputation and lost opportunities in the future. In addition to this problem of moral hazard (i.e., will agents behave cooperatively when completing economic transactions), reputation systems can address the problem of adverse selection (i.e., will high-quality agents choose to enter a market in the first place) (*76*, *77*).

A special challenge in an economy of AIs arises because of the fluidity of identity and the ease with which agents can be replaced. This raises, for example, the specter of whitewashing attacks, where an AI repeatedly runs down its reputation before reentering with a different identity. Without the possibility of enforcing strong identities that cannot be changed, this suggests a social cost of fluid identities, where it becomes necessary to impose a penalty on all new participants and make them build up reputations from an assumption of being untrustworthy (*78*).

We should also consider that *machina economicus* will be strategic in sharing feedback on other AIs. For example, in eBay's original reputation system, buyers were often reluctant to leave negative feedback about deadbeat sellers, because the sellers could retaliate with negative feedback about the buyer. In response, eBay introduced an additional feedback mechanism that was one-directional from the buyer to the seller and could not be easily traced to a particular buyer. The change resulted in a greater amount of negative feedback (*79*).

The economy of AIs also offers positive opportunities for promoting trust through bookkeeping, collecting feedback, and tracking the provenance of feedback in novel reputation mechanisms (see Fig. 5). AI researchers are designing reputation systems that align incentives with making truthful reports, while provably satisfying axiomatic properties such as symmetry: Two agents that are in an equivalent position from the perspective of reports made and received should have the same trust score (*80*, *81*). Another example is the design of accounting systems that elicit truthful reports about the resources contributed or work performed by other AIs and enable the design of systems to mitigate free-riding and promote fair contributions to an economic system (*82*). Still, the extent to which effective, multi-agent AIs can be developed entirely through computational infrastructure such as reputation mechanisms and without recourse to legal systems remains an interesting open question.

Closing comments

Whatever one's thoughts about when or whether AI will transcend human-level performance, the rapidly advancing capabilities of AI are fueling considerable optimism and investment in AI research. AI has surpassed or will likely soon surpass humans in narrow domains such as playing chess, controlling a jumbo jet during cruise, making product recommendations, pricing millions of products on an eCommerce platform, reasoning about whether a patient is likely to be re-admitted to a hospital, and detecting signals from a massive volume of financial news stories.

Certainly, many fundamental challenges remain, including how to design reasoning and inference methods that effectively balance the benefit of additional computation with the costs that may arise from additional delay to acting in the world and how to design AI systems that can learn and generalize from reward signals in unconstrained domains. Given that decision problems related to economic transactions are often relatively well structured, however, it seems likely to us that AI will continue to make especially rapid inroads in economically important applications. This in turn will ensure continued effort on methods for rational, economic reasoning toward the broader goal of developing *machina economicus*.

We should not leave the impression that AI researchers unanimously embrace economic perspectives on single- or multi-agent AI. For some, multi-agent economic models are still seen as a distraction. After all, a centralized perspective allows focusing on overall goals without worrying about the incentives of individual parts of the system. Others conduct research into multi-agent systems composed of agents under the control of the designer, so that they can be programmed in any way desired. Just as with centralized solutions, these so-called "cooperative" multi-agent systems allow design without concern for the self-interest of individual agents, albeit often with decomposition or communication constraints. But cooperative versus self-interested is really a difference in assumptions on the power of a system designer, rather than a technical dispute. The viewpoint that we ascribe to is that a large number of AI systems will, given the existing structure of human economic systems, be populated by AIs that are designed, deployed, owned, and operated by a myriad of different parties, each with possibly misaligned goals. Finally, some may object to the economic approach on the basis that AIs are and will remain far from perfectly rational, simply by virtue of physical and computational limits. More direct models of the AIs' computational behavior, in terms of the automata they are, could in principle be more accurate. The analytical utility of a rationality abstraction for AIs is ultimately an empirical question to be resolved as AI progresses.

Among those adopting an economic approach, there persist some disagreements on specific techniques—for example, on the role of equilibrium reasoning. Even if agents can be viewed as rational, some question whether it is plausible that they reach equilibrium configurations,

particularly in situations where multiple equilibria exist. As Shoham (*83*) argues, game theory lacks a well-accepted pragmatic account of how it should be deployed in concrete reasoning contexts. A positive view is that AI researchers, in their efforts to operationalize economic reasoning, are developing exactly this needed body of pragmatics.

Some may object that mechanism design is too idealized even for systems of AIs—for example, in its insistence on design under equilibrium behavior, its assumption that rules of interaction can be designed from scratch, and its lack of attention to the details of the human and legal contexts in which designed systems will operate. A positive view is that AI systems are precisely the kinds of environments where we can build *tabula rasa* new rules of interaction, because these rules will be realized through the Internet and as programs running on computer servers. That such rules of interaction can come into existence is as much a matter of science and engineering as it is of public policy.

As AI advances, we are confident that economic reasoning will continue to have an important role in the design of single-agent and multi-agent AIs, and we have argued that, as economies of AIs continue to emerge, there will need to be a new science to understand how to design these systems. These AIs will no doubt exert strong forces on the economy and broader society; understanding the effect and extent of this will shape the research agendas of both AI and economics in years to come.

REFERENCES AND NOTES

1. S. Russell, P. Norvig, *Artificial Intelligence: A Modern Approach* (Prentice Hall, ed. 3, 2009).
2. J. Doyle, *Comput. Intell.* **8**, 376–409 (1992).
3. E. J. Horvitz, Computation and action under bounded resources, thesis, Stanford University (1990).
4. S. Russell, in *Fundamental Issues of Artificial Intelligence*, V. C. Muller, Ed. (Springer, Berlin, 2015).
5. J. S. Rosenschein, G. Zlotkin, *Rules of Encounter: Designing Conventions for Automated Negotiation among Computers* (MIT Press, Cambridge, 1994).
6. M. McCue, Can't secure a reservation to a hot restaurant? Try using an algorithm, Fortune, 3 September 2014; http://fortune.com/2014/09/03/opentable-dinematic-tablesweep-restaurant-reservation-service/.
7. J. D. Sutter, CNN, 25 April 2011; http://edition.cnn.com/2011/TECH/web/04/25/amazon.price.algorithm.
8. T. Hendershott, C. M. Jones, A. J. Menkveld, *J. Finance* **66**, 1–33 (2011).
9. D. J. Abraham, A. Blum, T. Sandholm, 8th ACM Conference on Electronic Commerce (2007), pp. 295–304.
10. A. Frechette, N. Newman, K. Leyton-Brown, 24th International Joint Conference on Artificial Intelligence (2015).
11. S. D. Ramchurn, P. Vytelingum, A. Rogers, N. R. Jennings, *Commun. ACM* **55**, 86–97 (2012).
12. R. I. Brafman, M. Tennenholtz, *Artif. Intell.* **159**, 27–47 (2004).
13. A. X. Jiang *et al*., *Adv. Neural Inf. Process. Syst.* **27**, 2573–2581 (2014).
14. A. Y. Ng, D. Harada, S. J. Russell, 16th International Conference on Machine Learning (1999), pp. 278–287.
15. S. Singh, R. L. Lewis, A. G. Barto, J. Sorg, *IEEE Transactions on Autonomous Mental Development* **2**, 70–82 (2010).
16. L. Getoor, B. Taskar, Eds., *Introduction to Statistical Relational Learning* (MIT Press, Cambridge, 2007).
17. M. I. Jordan, T. M. Mitchell, *Science* **349**, 255–260 (2015).
18. J. Hirschberg, C. D. Manning, *Science* **349**, 261–266 (2015).
19. Y. Shoham, K. Leyton-Brown, Multiagent Systems: Algorithmic, Game-Theoretic and Logical Foundations (Cambridge Univ. Press, Cambridge, 2009).
20. D. Fudenberg, D. K. Levine, *The Theory of Learning in Games* (MIT Press, Cambridge, 1998).
21. M. P. Wellman, D. M. Reeves, K. M. Lochner, Y. Vorobeychik, *J. Artif. Intell. Res.* **21**, 19–36 (2004).
22. D. Billings, A. Davidson, J. Schaeffer, D. Szafron, *Artif. Intell.* **134**, 201–240 (2002).
23. M. Bowling, N. Burch, M. Johanson, O. Tammelin, *Science* **347**, 145–149 (2015).
24. T. Sandholm, *AI Mag.* **31**, 13–32 (2010).
25. N. Brown, S. Ganzfried, T. Sandholm, 14th International Conference on Autonomous Agents and Multi-Agent Systems (2015), pp. 7–15.
26. R. Gibson, Regret minimization in games and the development of champion multiplayer computer poker-playing agents, Ph.D. thesis, University of Alberta (2014).
27. M. Tambe, *Security and Game Theory: Algorithms, Deployed Systems, Lessons Learned* (Cambridge Univ. Press, Cambridge, 2011).
28. Y. Gal, A. Pfeffer, *J. Artif. Intell. Res.* **33**, 109–147 (2008).
29. C. F. Camerer, *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, Princeton, 2003).
30. R. Yang, C. Kiekintveld, F. Ordóñez, M. Tambe, R. John, *Artif. Intell.* **195**, 440–469 (2013).
31. J. R. Wright, K. Leyton-Brown, Evaluating, understanding, and improving behavioral game theory models for predicting human behavior in unrepeated normal-form games, CoRR abs/1306.0918 (2013); http://arxiv.org/abs/1306.0918.
32. D. Monderer, M. Tennenholtz, *Artif. Intell.* **173**, 180–195 (2009).
33. L. Hurwicz, in *Mathematical Methods in the Social Sciences*, K. J. Arrow, S. Karlin, P. Suppes, Eds. (Stanford University Press, Stanford 1960), Ch. 3, pp. 27–46.
34. T. Sönmez, U. Ünver, in *Handbook of Social Economics*, A. Bisin, J. Benhabib, M. Jackson, Eds. (North-Holland, 2011), vol. 1A, pp. 781–852.
35. P. Milgrom, *Putting Auction Theory to Work* (Cambridge Univ. Press, Cambridge, 2004).
36. H. R. Varian, 1st USENIX Workshop on Electronic Commerce (1995), pp. 13–21.
37. W. Vickrey, *J. Finance* **16**, 8–37 (1961).
38. Y. Chen, T. Sönmez, *J. Econ. Theory* **127**, 202–231 (2006).
39. F. Echenique, A. J. Wilson, L. Yariv, Clearinghouses for two-sided matching: An experimental study, Working Papers 487, University of Pittsburgh, Department of Economics (2013).
40. N. Nisan, in Combinatorial Auctions, P. Cramton, Y. Shoham, R. Steinberg, Eds. (MIT Press, Cambridge, 2006), chap. 9.
41. T. Sandholm, C. Boutilier, in Combinatorial Auctions, P. Cramton, Y. Shoham, R. Steinberg, Eds. (MIT Press, Cambridge, 2006), chap. 10.
42. T. Sandholm, *Artif. Intell.* **135**, 1–54 (2002).
43. V. Conitzer, T. Sandholm, 18th Conference on Uncertainty in Artificial Intelligence (2002), pp. 103–110.
44. S. Alaei, H. Fu, N. Haghpanah, J. D. Hartline, A. Malekian, 13th ACM Conference on Electronic Commerce (2012), p. 17.
45. Y. Cai, C. Daskalakis, S. M. Weinberg, 54th Annual IEEE Symposium on Foundations of Computer Science (2013), pp. 618–627.
46. P. Duetting *et al*., *ACM Transactions on Economics and Computation* **3**, 5:1-5:41 (2015).
47. D. C. Parkes, S. Singh, *Adv. Neural Inf. Process. Syst.* **16**, 791–798 (2003).
48. R. Cavallo, D. C. Parkes, S. Singh, 22nd Conference on Uncertainty in Artificial Intelligence (Cambridge, MA, 2006), pp. 55–62.
49. D. C. Parkes, in *Algorithmic Game Theory*, N. Nisan, T. Roughgarden, E. Tardos, V. Vazirani, Eds. (Cambridge Univ. Press, 2007), chap. 16, pp. 411–439.
50. Y. Shoham, M. Tennenholtz, *Theor. Comput. Sci.* **343**, 97–113 (2005).
51. J. Y. Halpern, V. Teague, 36th Annual ACM Symposium on Theory of Computing (2004), pp. 623–632.
52. N. Alon, F. Fischer, A. Procaccia, M. Tennenholtz, 13th Conference on Theoretical Aspects of Rationality and Knowledge (2011), pp. 101–110.
53. O. Dekel, F. A. Fischer, A. D. Procaccia, *J. Comput. Syst. Sci.* **76**, 759–777 (2010).
54. M. Yokoo, Y. Sakurai, S. Matsubara, *Games Econ. Behav.* **46**, 174–188 (2004).
55. B. Edelman, M. Ostrovsky, *Decis. Support Syst.* **43**, 192–198 (2007).
56. B. Edelman, M. Ostrovsky, M. Schwarz, *Am. Econ. Rev.* **97**, 242–259 (2007).
57. C. Borgs *et al*., 16th International Conference on World-Wide Web (2007), pp. 531–540.
58. J. Feldman, S. Muthukrishnan, M. Pál, C. Stein, 8th ACM Conference on Electronic Commerce (2007), pp. 40–49.
59. A. Z. Broder, E. Gabrilovich, V. Josifovski, G. Mavromatis, A. J. Smola, 4th International Conference on Web Search and Web Data Mining (2011), pp. 515–524.
60. L. Hurwicz, in *Decision and Organization*, R. Radner, C. B. McGuire, Eds. (North-Holland, Amsterdam, 1972), Ch. 14, pp. 297–336.
61. A. Gibbard, *Econometrica* **41**, 587–602 (1973).
62. R. Myerson, *Econometrica* **47**, 61–73 (1979).
63. E. Clarke, *Public Choice* **11**, 17–33 (1971).
64. T. Groves, *Econometrica* **41**, 617–631 (1973).
65. H. R. Varian, C. Harris, *Am. Econ. Rev.* **104**, 442–445 (2014).
66. M. Kearns, Y. Nevmyvaka, in *High Frequency Trading: New Realities for Traders, Markets and Regulators*, D. Easley, M. Lopez de Prado, M. O'Hara, Eds. (Risk Books, London, 2013), Ch. 5, pp. 91–124.
67. E. Budish, P. Cramton, J. Shim, The high-frequency trading arms race: Frequent batch auctions as a market design response, Tech. Rep. 14-03, Booth School of Business, University of Chicago (2015).
68. E. Wah, M. P. Wellman, 14th ACM Conference on Electronic Commerce (2013), pp. 855–872.
69. R. Forsythe, F. Nelson, G. R. Neumann, J. Wright, in *Contemporary Laboratory Experiments in Political Economy*, T. R. Palfrey, Ed. (University of Michigan Press, Ann Arbor, 1991), pp. 69–111.
70. R. D. Hanson, *Inf. Syst. Front.* **5**, 107–119 (2003).
71. J. Abernethy, Y. Chen, J. W. Vaughan, *ACM Transactions on Economics and Computation* **1**, 12:1-12:39 (2013).
72. M. Dudík, S. Lahaie, D. M. Pennock, D. Rothschild, 14th ACM Conference on Electronic Commerce (2013), pp. 341–358.
73. J. Abernethy, R. Frongillo, *Adv. Neural Inf. Process. Syst.* **24**, 2600–2608 (2011).
74. J. Witkowski, D. C. Parkes, 13th ACM Conference on Electronic Commerce (2012), pp. 964–981.
75. J. Hu, A. J. Storkey, 31st International Conference on Machine Learning (2014), pp. 1773– 1781.
76. P. Resnick, K. Kuwabara, R. Zeckhauser, E. J. Friedman, *Commun. ACM* **43**, 45–48 (2000).
77. C. Dellarocas, *Manage. Sci.* **49**, 1407–1424 (2003).
78. E. J. Friedman, P. Resnick, *J. Econ. Manage. Strategy* **10**, 173–199 (2001).
79. G. Bolton, B. Greiner, A. Ockenfels, *Manage. Sci.* **59**, 265–285 (2013).
80. A. Cheng, E. Friedman, ACM SIGCOMM Workshop on Economics of Peer-to-Peer Systems (2005), pp. 128–132.
81. A. Altman, M. Tennenholtz, *J. Artif. Intell. Res.* **31**, 473–495 (2008).
82. S. Seuken, J. Tang, D. C. Parkes, 24th AAAI Conference on Artificial Intelligence (2010), pp. 860–866.
83. Y. Shoham, *Commun. ACM* **51**, 74–79 (2008).
84. J. Sorg, S. Singh, R. Lewis, 27th International Conference on Machine Learning (2010), pp. 1007–1014.
85. D. Koller, A. Pfeffer, *Artif. Intell.* **94**, 167–215 (1997).
86. D. Billings *et al*., *18th International Joint Conference on Artificial Intelligence* (2003), pp. 661–668.
87. A. Gilpin, T. Sandholm, *7th ACM Conference on Electronic Commerce* (2006), pp. 160–169.
88. M. Zinkevich, M. Johanson, M. Bowling, C. Piccione, *Adv. Neural Inf. Process. Syst.* **20**, 905–912 (2007).
89. A. Gilpin, S. Hoda, J. Peña, T. Sandholm, *3rd International Workshop on Internet and Network Economics* (2007), pp. 57–69.
90. S. Hoda, A. Gilpin, J. Peña, T. Sandholm, *Math. Oper. Res.* **35**, 494–512 (2010).

ACKNOWLEDGMENTS

Thanks to the anonymous referees and to M. Bowling, Y. Chen, K. Gal, S. Lahaie, D. Pennock, A. Procaccia, T. Sandholm, S. Seuken, Y. Shoham, A. Storkey, M. Tambe, and M. Tennenholtz for their thoughtful comments on an earlier draft.

10.1126/science.aaa8403