Oxford Review of Economic Policy, Volume 37, Number 3, 2021, pp. 509–520

The history and future of AI

Stuart Russell*

Abstract: The standard model for developing AI systems assumes a fixed, known objective that the AI system is required to optimize through its actions. Systems developed within the standard model have been increasingly successful. I briefly summarize the state of the art and its likely evolution over the next decade. Substantial breakthroughs leading to general-purpose AI are much harder to predict, but they will have an enormous impact on the global economy and on human roles therein. At the same time, I expect that the standard model will become increasingly untenable in real-world applications because of the difficulty of specifying objectives completely and correctly. I propose a new model for AI development in which the machine's uncertainty about the true objective leads to qualitatively new modes of behaviour that are more robust, controllable, and deferential.

Keywords: artificial intelligence, rationality, machine learning, future of work

JEL classification: N10, O31, O32, O47

I. Introduction

Developments in artificial intelligence (AI) are generating significant media interest and commercial investment. AI has become the focus of great-power rivalry, with many commentators viewing it as the technological underpinning of future economic ascendancy. Others view the present level of excitement as yet another 'hype cycle', repeating the overenthusiasm of the 1960s and 1980s. A more defensible view distinguishes cumulative advances in research from public demonstrations and commercial exploitation.

The purpose of this paper is to situate the present capabilities of AI within a longer intellectual tradition, the better to anticipate likely future developments. That tradition, in which economic theory has played a significant role, equates intelligence with the ability to act rationally—that is, to choose actions that can be expected to achieve one's objectives. This framework, explicated in section II of the paper, is so pervasive within AI that it would be reasonable to call it the standard model. A great deal of progress on reasoning, planning, and decision making, as well as perception and learning, has occurred within the standard model. As discussed in section III, these advances are likely to lead within the next decade to new, practical capabilities of significant economic value.

In section IV, I discuss the possibility of what is variously called human-level AI, general-purpose AI, or artificial general intelligence (AGI). Although we are far from

*Computer Science Division, University of California, Berkeley, USA, e-mail: russell@cs.berkeley.edu

The research reported herein was supported by grants from the Future of Life Institute, the Leverhulme Trust, and the Open Philanthropy Foundation.

doi:10.1093/oxrep/grab013

achieving this goal, it is reasonable to suppose that eventual success would have economic impacts so far-reaching as to be almost incalculable. At the same time, we cannot ignore the social consequences of replacing humans in the vast majority of currently valued economic roles. Anticipating and preparing for desirable socioeconomic arrangements in this new era is an important task for economists, policy-makers, and a wide range of academic disciplines.

Section V addresses a different consequence of progress in AI: the potential for loss of control over increasingly capable AI systems. In 1951, Alan Turing spoke on the BBC Third Programme as follows: 'It seems probable that once the machine thinking method had started, it would not take long to outstrip our feeble powers. . . . At some stage therefore we should have to expect the machines to take control.' Unlike Turing, I believe this fate is avoidable, but it means abandoning the standard model in favour of one in which machines are necessarily beneficial rather than merely intelligent. This new model has interesting similarities to the principal–agent models studied in economics and draws on the tools of preference elicitation, game theory, and mechanism design.

II. Intelligence as rationality and the standard model of AI

The central technical concept in AI is that of an *agent*—an entity that perceives and acts (Russell and Norvig, 2020).[1] Cognitive faculties such as reasoning, planning, and learning are in the service of acting. The concept can be applied to humans, robots, software entities, corporations, nations, or thermostats. AI is concerned principally with designing the internals of the agent: mapping from a stream of raw perceptual data to a stream of actions. Designs for AI systems vary enormously depending on the nature of the environment in which the system will operate, the nature of the perceptual and motor connections between agent and environment, and the requirements of the task.

AI seeks agent designs that exhibit 'intelligence', but what does this mean? Aristotle (in his *Ethics*) gave one answer: 'We deliberate not about ends, but about means. . . . [We] assume the end and consider how and by what means it is attained, and if it seems easily and best produced thereby.' That is, an intelligent or *rational* action is one that can be expected to achieve one's objectives. This line of thinking has persisted to the present day. Arnauld (1662) broadened Aristotle's theory to include uncertainty in a quantitative way, proposing that we should act to maximize the *expected value* of the outcome. Daniel Bernoulli (1738) refined the notion of value, moving it from an external quantity (typically money) to an internal quantity that he called *utility*. De Montmort (1713) noted that in games (decision situations involving two or more agents) a rational agent might have to act randomly to avoid being second-guessed. Von Neumann and Morgenstern (1944) tied all these ideas together into an axiomatic framework that underlies much of modern economic theory.

As AI emerged in the 1940s and 1950s, it needed some notion of intelligence on which to build the foundations of the field. Although some early research was aimed more at emulating human cognition, the notion that won out was rationality: a machine is intelligent to the extent that its actions can be expected to achieve its objectives. In the

[1] The word 'agent' in AI carries no connotation of acting on behalf of another.

standard model, we aim to build machines of this kind; we define the objectives; and the machine does the rest. There are several different ways in which the standard model can be instantiated. For example, a problem-solving system for a deterministic environment is given a cost function and a goal criterion and finds the least-cost action sequence that leads to a goal state; a reinforcement learning system for a stochastic environment is given a reward function and a discount factor and learns a policy that maximizes the expected discounted sum of rewards.

This general approach is not unique to AI. Control theorists minimize cost functions; operations researchers maximize rewards; statisticians minimize an expected loss function; and economists, of course, maximize the utility of individuals, the welfare of groups, or the profit of corporations.

III. Capabilities and trends

Among the many myths about AI that circulate in the media, perhaps the most misleading is the idea that AI is a new technology, sometimes called 'deep learning', that emerged in the mid-2010s as a result of 'breakthroughs' such as the victory of AlphaGo over Lee Sedol, the former human Go world champion. In fact, the current capabilities of AI systems derive from decades of research and development in a variety of areas (of which deep learning is just one), building on centuries of work on formal models of thinking and acting. What the media call 'breakthroughs'—such as victories by DeepBlue in chess, Watson in Jeopardy!, and AlphaGo in Go—are more properly called 'demonstrations'. In all three cases, these demonstrations showcased technology derived from multiple research breakthroughs that occurred decades earlier.

Perhaps the oldest-established area of AI is that of combinatorial search, in which algorithms consider many possible sequences of future actions or many possible configurations of complex objects. Examples include route-finding algorithms for GPS navigation, robot assembly planning, transportation scheduling, and protein design. Closely related algorithms are used in game-playing systems. In all of these algorithms, the key issue is efficient exploration to find good solutions quickly, despite the vast search spaces inherent in combinatorial problems.

Beginning in around 1960, AI researchers and mathematical logicians developed ways to represent logical assertions as data structures as well as algorithms for performing logical inference with those assertions. Since that time, the technology of so-called automated reasoning has advanced dramatically. For example, it is now routine to verify the correctness of VLSI designs before production and the correctness of software systems and cybersecurity protocols before deployment in high-stakes applications. The technology of logic programming (and related methods in database systems) makes it easy to specify and check the application of complex sets of logical rules in areas such as insurance claims processing, data system maintenance, security access control, tax calculations, and government benefit distribution. Special-purpose reasoning systems designed to reason about actions can construct large-scale, provably correct plans in areas such as logistics, construction, and manufacturing. The most visible application of logic-based representation and reasoning is Google's Knowledge Graph, which, as of May 2020, holds 500 billion facts about 5 billion entities (Sullivan,

2020) and is used to answer directly more than a third of all queries submitted to the Google search engine.

Beginning in the 1980s, the AI community began to grapple with the uncertainty inherent in real-world observations and in knowledge acquired from humans or through machine learning. Although some rule-based expert systems adopted *ad hoc* calculi for representing and propagating uncertainty, probability theory became the dominant tool, largely due to the development of Bayesian networks by Judea Pearl (1988) and others. This led to the development of the first large-scale computational tools for probabilistic reasoning and to substantial cross-fertilization between AI and other fields that build on probability theory, including statistics, information theory, control theory, and operations research. Bayesian networks and related methods have been used for modelling, diagnosis, monitoring, and prediction of a wide range of complex systems including jet engines, Mars rovers, ecological networks, and intensive care patients. Causal networks (Pearl, 2000; Pearl and Mackenzie, 2018), which extend Bayesian networks to model the effects of exogenous interventions, have clarified and facilitated the analysis of causal relationships in many empirical disciplines, especially in the social sciences.

The development of probabilistic programming languages or PPLs (Koller *et al.*, 1997; Pfeffer, 2001; Milch *et al.*, 2005; Goodman *et al.*, 2008) provides a *universal* representation for probability models, meaning that any model representable in any formalism can be represented efficiently in a PPL. Moreover, PPLs come with general-purpose inference algorithms, so that (in principle, at least) no algorithm development or mathematical derivations are needed when applying probability theory to a new domain. PPLs constitute one of the fastest-growing areas of AI and enable the rapid construction of enormously complex models. For example, the new monitoring system for the Comprehensive Nuclear-Test-Ban Treaty began life as a PPL model that took only a few minutes to write (Le Bras *et al.*, 2020); while operating, it may dynamically construct internal representations involving hundreds of thousands of random variables.

Alan Turing (1950) suggested that machine learning would be the most practical way to create AI capabilities. The most common paradigm is *supervised learning*, wherein labelled examples are provided to a learning algorithm that outputs a predictive hypothesis with which to label unlabelled examples. Early developments in AI and in statistics proceeded separately, but both fields produced useful tools for learning low-dimensional models, with application to areas such as loan decisions, credit-card fraud detection, and email spam filtering. For high-dimensional data such as images, deep convolutional networks have proved to be effective (LeCun *et al.*, 1989, 2015; Krizhevsky *et al.*, 2013). Deep learning has substantially advanced the state of the art in visual object recognition, speech recognition, and machine translation, three of the most important subfields of AI, as well as in protein folding, a key problem in molecular biology. Language models such as GPT-3—very large neural networks trained to predict the next word in a sequence—show intriguing abilities to respond to questions in a semantically meaningful way. Recent work has shown, however, that deep learning systems often fail to generalize robustly and are susceptible to spurious regularities in the training data (Carter *et al.*, 2020; D'Amour *et al.*, 2020). Moreover, the amount of training data required to achieve a given level of performance is far greater than a human typically requires.

The algorithmic study of sequential decision-making under uncertainty began in economics (Shapley, 1953) and operations research (Bellman, 1952, 1957). Algorithms

developed in these fields typically handle only small problems with up to a million states. In AI, the development of *reinforcement learning* (RL) has allowed much larger problems to be addressed satisfactorily, including checkers with 10^{21} positions (Samuel, 1959) and Go with 10^{170} positions (Silver *et al.*, 2016). RL algorithms learn by experiencing state transitions and their associated rewards while updating a representation of the value of states (and possibly actions as well) or a direct representation of the decision policy. Applications of RL range from bidding in advertising markets (Jin *et al.*, 2018) to improving the ability of robots to grasp previously unseen objects (Quillen *et al.*, 2018). As with supervised learning, applications of deep networks in RL may also be quite fragile (Gleave *et al.*, 2020).

With modest advances in perception and dexterity, we can expect to see robots moving into a range of unstructured environments including roads, warehouses, agriculture, mining, and warfare. We may see progress on language understanding comparable to the progress on image understanding that occurred in the last decade, which would enable high-impact applications such as intelligent personal assistants and high-quality intelligent tutoring systems. Search engines, rather than responding to keywords with URLs, would respond to questions with answers based on reading and, to some degree, understanding everything the human race has ever written. And text would be augmented by satellite imagery, enabling computers to see every object (50cm or larger) on Earth every day.

Although this view is far from universally shared, I think it is likely that in the coming decade the pendulum will swing away from a reliance on end-to-end deep learning and back towards systems composed from modular, semantically well-defined representations built on the mathematical foundations of logic and probability theory, with deep learning playing a crucial role in connecting to raw perceptual data. (This approach underlies, for example, Waymo's industry-leading self-driving car project.) The reasons for this prediction are complex, but include (i) the performance problems with deep learning mentioned earlier; (ii) the advantages, in terms of rigor and transparency, of being able to analyse systems as *possessing knowledge* and *reasoning* with that knowledge; (iii) the expressive limitations of circuit-based representations (including deep learning systems) for capturing general knowledge; (iv) the essential role played by prior knowledge in enabling a learning system to generalize robustly from small numbers of examples; and (v) the enormous benefits of being able to improve the performance of systems by supplying knowledge rather than training data. It is important to understand that 'modular, semantically well-defined representations' are not necessarily hand-engineered or inflexible: such representations can be learned from data, just as the entire edifice of science itself is a modular, semantically well-defined representation that has (ultimately) been learned from data.

IV. Future developments and implications

The existing and projected advances described in the preceding section will undoubtedly lead to economically important applications in many areas, and many forecasters see these developments as contributing significantly to economic growth over the next decade. Yet this effect is likely to be small compared to that of truly general-purpose AI.

General-purpose AI has been the long-term goal of the field since its inception. It would be as flexible as human intelligence: given essentially any task that is feasible for a human or collection of humans, it would quickly learn to perform the task as well as or better than humans. It would also have massive speed, memory, and input bandwidth advantages over humans. Its potential benefits would be far greater than those of a collection of narrow, application-specific AI systems, just as supplying general-purpose electrical power has had far more benefits than the varied and sometimes bizarre collection of electrostatic phenomena and devices so ingeniously contrived by eighteenth-century physicists. For this reason, the prospect of creating general-purpose AI is driving massive investments and geopolitical rivalries.

We are far from achieving general-purpose AI.[2] No amount of data or computing power is going to change that. We need conceptual breakthroughs in a number of areas including decision-making over long timescales and the cumulative use of knowledge in learning. These breakthroughs are inherently unpredictable. In a 1977 interview, John McCarthy, one of the 'founding fathers' of AI, said, 'What you want is 1.7 Einsteins and 0.3 of the Manhattan Project, and you want the Einsteins first. I believe it'll take five to 500 years' (Shenker, 1977). This remains true today, although we have seen dramatic progress since 1977 in many areas. The vast majority of AI researchers now believe that general-purpose, human-level AI will arrive in this century (Grace *et al.*, 2018).

It is important to understand, moreover, that most of the progress towards general-purpose AI has occurred as a result of working on narrow, special-purpose applications. This is because AI researchers working on such applications are typically developing not merely an *ad hoc* encoding of what an intelligent person would do in such-and-such situation, but an attempt to provide the machine with the ability to figure out the solution for itself. For example, when Yann LeCun's team at AT&T Labs worked on recognizing handwritten digits in the 1990s, they didn't write special algorithms to recognize '8' by searching for curvy lines and loops; instead, they improved on existing neural network learning algorithms to produce *convolutional* neural networks. Those networks can also learn to recognize letters, shapes, stop signs, dogs, cats, police cars, and desirable Go positions; in fact, they were the basis for the deep-learning revolution of the 2010s.

The consequences of developing general-purpose AI technology are significant. One can speculate about solving major open problems, such as extending human life indefinitely or developing faster-than-light travel, but these staples of science fiction are not yet the driving force for progress in AI. Consider, instead, a more prosaic goal: raising the living standard of everyone on Earth, in a sustainable way, to a level that would be viewed as quite respectable in a developed country. Choosing (somewhat arbitrarily) respectable to mean the eighty-eighth percentile in the United States, the stated goal represents almost a tenfold increase in global gross domestic product (GDP), from $76 trillion to $750 trillion per year. The net present value of the increased income stream is $13.5 quadrillion, assuming a discount factor of 5 per cent.[3] (The value is $9.4

[2] There would be no single moment at which AI exceeds human intelligence. AI capabilities in different branches of cognitive activity vary far more than those of humans, and in some branches they already exceed human capacities by orders of magnitude. (For example, a search engine remembers very well and cannot plan at all; a chess program plans very well and cannot remember at all.) By the time that AI systems exhibit generality across all branches, direct comparisons to humans will be meaningless.

[3] I. J. Good, who coined the phrase 'intelligence explosion', estimated the value of general-purpose AI at one megaKeynes, equivalent to around $2,500 quadrillion in current dollars.

quadrillion or $6.8 quadrillion if the technology is phased in over 10 or 20 years.) These numbers are large relative to the amounts currently invested in AI research.

A tenfold increase in global GDP *per capita* took place over 190 years, from 1820 to 2010 (van Zanden, 2014). Achieving that tenfold increase required the development of factories, machine tools, automation, railways, steel, cars, airplanes, electricity, oil and gas production, telephones, radio, television, computers, the Internet, satellites, and many other revolutionary inventions. The tenfold increase in GDP posited above is predicated not on further revolutionary technologies but on the ability of general-purpose AI systems to employ what we already have more effectively and at greater scale. There would be no need to employ armies of specialists in different disciplines, organized into hierarchies of contractors and subcontractors, in order to carry out a project. All embodiments of general-purpose AI would have access to all the knowledge and skills of the human race, and more besides. The only differentiation would be in the physical capabilities: dexterous legged robots for construction or surgery, wheeled robots for large-scale goods transportation, quadcopter robots for aerial inspections, and so on. In principle—politics and economics aside[4]—everyone could have at their disposal an entire organization composed of software agents and physical robots, capable of designing and building bridges or (fully automated) factories, improving crop yields, cooking dinner for a hundred guests, running elections, teaching a child to read, or doing whatever else needs doing. It is the generality of general-purpose intelligence that makes this possible.

In this role as a universal producer of goods and services, general-purpose AI will have an effect on the global economy that is quite different from that of other general-purpose technologies such as electrical power. With the possible exceptions of heat and light, electrical power did not directly meet human needs, but it did improve many existing supply chains for meeting human needs and it helped bring new supply chains into existence (such as those based on computation). General-purpose AI *also* has these characteristics—for example, as a tool that can improve the quality and speed of scientific research, the allocation of scarce resources, and the coordination of human activities.

The advent of general-purpose AI will inevitably cause far-reaching changes in the organization and productivity of our economic systems, but it will also affect humans' roles therein. Aristotle, in Book I of his *Politics*, first described what Keynes (1930) later called *technological unemployment*:

> For if every instrument could accomplish its own work, obeying or anticipating the will of others . . . if, in like manner, the shuttle would weave and the plectrum touch the lyre without a hand to guide them, chief workmen would not want servants, nor masters slaves.

In the last decade, thousands of media articles and opinion pieces and many books have been written on this topic (e.g. Ford, 2015; Chace, 2016; Brynjolfsson and McAfee, 2016;

[4] The political and economic difficulties should not, of course, be underestimated. Corporations, elites, or countries may attempt to hoard general-purpose AI technology and its benefits, and under some circumstances economic incentives may operate to retard the dissemination of AI-based goods and services (Aghion *et al.*, 2017). One can also expect finite resources such as land, human attention, and perhaps raw materials to become relatively more expensive.

Agrawal *et al.*, 2019). Others well-known economists sounding the alarm include Robert Shiller, Michael Spence, Paul Krugman, Klaus Schwab, and Larry Summers. Research centres are springing up all over the world to understand what is likely to happen. Will new jobs arise to replace all those done by the machines, or will machines do the new jobs too? Put another way: when a machine replaces one's physical labour, one can sell mental labour. When a machine replaces one's mental labour, what does one have left to sell?

Bessen (2019) posits a simple quantitative model whereby increased productivity through automation initially reduce prices, raises demand, and *increases* employment in a given industry. Eventually, demand saturates and further technological advances lead to reduced employment. Over time, the employment numbers describe an 'inverted-U' curve. (See also Russell (2019, pp. 113–24).) Bessen provides such curves for several industries in the twentieth century and makes the obvious point that the effect of a technological advance on employment depends on which side of the curve that industry is on at the time. With general-purpose AI, every industry will move quickly to the downhill side of the inverted U.

One possible response is the provision of a universal basic income (UBI). Perhaps surprisingly, UBI has support across the political spectrum, ranging from the Adam Smith Institute (Bowman, 2013) to the Green Party (Bartley, 2017). Among Voltaire's three great evils, UBI addresses 'need' but not 'boredom' and 'vice'. Future economic roles for humans will be ones in which we simply *prefer* to have humans, or where humans have an intrinsic competitive advantage because the similarity of our nervous systems and cognitive architectures enables us to know *what it's like* for another human to have a given experience. Such roles are likely to be in interpersonal services to individuals.

At present, most such roles—such as childcare—are low-status and poorly remunerated. This is not because we do not value children, but because we lack the necessary science base and professional training curricula to deliver childcare that is highly beneficial and therefore of high economic value. (Compare childcare with, for example, cardiac care.) The same is true for essentially all avenues by which one person can improve the life of another by direct interaction. The required level of scientific understanding of human psychology and human life may take decades to create.

V. Failure of the standard model and a possible replacement

As described in section II, the standard model of AI (and related disciplines) is a pillar of twentieth-century technology. Unfortunately, this standard model is a mistake. Once AI systems move out of the laboratory (or artificially defined environments such as the simulated Go board) and into the real world, there is very little chance that we can specify our objectives completely and correctly in such a way that the pursuit of those objectives by more capable machines is guaranteed to result in beneficial outcomes for humans. Indeed, we may lose control altogether, as noted by Turing in his 1951 lecture, as machines take pre-emptive steps to ensure that the stated objective is achieved.[5] This

[5] Hillis (2019), among others, has drawn the analogy between uncontrollable AI systems and uncontrollable economic actors—such as fossil-fuel corporations maximizing profit at the expense of humanity's future. Machine learning algorithms performing content selection on social media platforms may be the most significant illustration so far of the risks posed by AI systems pursuing incorrectly defined objectives; indeed, such algorithms optimize their reward functions by manipulating humans to make them more predictable in their clicking behaviour (Groth *et al.*, 2019).

is why, when a genie has granted three wishes, the third wish is always to undo the first two wishes.

The mistake comes from transferring a perfectly reasonable definition of intelligence from humans to machines. The definition is reasonable for humans because we are entitled to pursue our own objectives. (Indeed, whose would we pursue, if not our own?) Machines, on the other hand, are not entitled to pursue their own objectives. A more sensible definition of AI would have machines pursuing *our* objectives. In the unlikely event that we can specify the objectives completely and correctly and insert them into the machine, then we can recover the standard model as a special case. If not, then the machine will necessarily be uncertain as to our objectives, while being obliged to pursue them on our behalf. This uncertainty—with the coupling between machines and humans that it entails—is crucial to building AI systems of arbitrary intelligence that are provably beneficial to humans.

In the 1980s the AI community abandoned the idea that AI systems could have definite knowledge of the state of the world or of the effects of actions, and it embraced uncertainty in these aspects of the problem statement. It is not at all clear why, for the most part, the community failed to notice that there must also be uncertainty in the objective. Although some AI problems, such as puzzle solving, have inherently well-defined goals, many other problems that were considered at the time, such as recommending medical treatments, have no precise objectives and ought to reflect the fact that the relevant preferences (of patients, relatives, doctors, insurers, hospital systems, taxpayers, etc.) are not known initially in each case. While it is true that unresolvable uncertainty over objectives can be integrated out of any decision problem, leaving an equivalent decision problem with a definite (average) objective, this transformation is invalid when there is the possibility of additional evidence regarding the true objectives. Thus, one may characterize the primary difference between the standard and new models of AI through the flow of preference information from humans to machines at 'run-time'. This flow comes from evidence provided by human behaviour.

This basic idea is made more precise in the framework of assistance games—originally known as cooperative inverse reinforcement learning (CIRL) games in the terminology of Hadfield-Menell *et al.* (2017*a*). The simplest case of an assistance game involves two agents, one human and the other a robot. It is a game of partial information, because, while the human knows the reward function, the robot does not——even though the robot's job is to maximize it. In a Bayesian formulation, the robot begins with a prior probability distribution over the human reward function and updates it as the robot and human interact during the game. Assistance games can be generalized to allow for imperfectly rational humans (Hadfield-Menell *et al.*, 2017*b*), humans who don't know their own preferences (Chan *et al.*, 2019), multiple human participants (Fickinger *et al.*, 2020), multiple robots, and so on.

Assistance games are connected to inverse reinforcement learning or IRL (Russell, 1998; Ng and Russell, 2000) because the robot can learn more about human preferences from the observation of human behaviour——a process that is the dual of reinforcement learning, wherein behaviour is learned from rewards and punishments. The primary difference is that in the assistance game, unlike the IRL framework, the human's actions are affected by the robot's presence—for example, the human may try to teach the robot about his or her preferences.

The overall approach also resembles principal–agent problems in economics, wherein the principal (e.g. an employer) needs to incentivize another agent (e.g. an employee) to behave in ways beneficial to the principal. The key difference here is that, unlike a human

employee, the robot has no interests of its own. Furthermore, we are building one of the agents in order to benefit the other, so the appropriate solution concepts may differ.

Within the framework of assistance games, a number of basic results can be established that are relevant to Turing's problem of control.

- Under certain assumptions about the support and bias of the robot's prior over human rewards, one can show that a robot solving an assistance game has non-negative value to humans (Hadfield-Menell *et al.*, 2017*a*).
- A robot that is uncertain about the human's preferences has a non-negative incentive to allow itself to be switched off (Hadfield-Menell *et al.*, 2017*b*). In general, it will defer to human control actions.
- To avoid changing attributes of the world whose value is unknown, the robot will generally engage in 'minimally invasive' behaviour to benefit the human (Shah *et al.*, 2019). Even when it knows nothing at all about human preferences, it will still take 'empowering' actions that expand the set of actions available to the human.

There are too many open research problems in the new model of AI to list them all here. The one most directly relevant to economics and public policy is the question of *social aggregation*: how should a machine make decisions when its actions affect the interests of more than one human being? Obviously, this question is central to moral philosophy and the social sciences. Issues include the preferences of evil individuals (Harsanyi, 1977); relative preferences and positional goods (Veblen, 1899; Hirsch, 1977); and interpersonal comparison of preferences (Nozick, 1974; Sen, 1999). Also of great importance is the *plasticity* of human preferences, which brings up both the philosophical problem of how to decide on behalf of a human whose preferences change over time (Pettigrew, 2020) and the practical problem of how to ensure that AI systems are not incentivized to change human preferences in order to make them easier to satisfy.

Assuming that the theoretical and algorithmic foundations of the new model for AI can be completed and then instantiated in the form of useful systems such as personal digital assistants or household robots, it will be necessary to create a technical consensus around a set of design templates for provably beneficial AI, so that policymakers have some concrete guidance on what sorts of regulations might make sense. The economic incentives would tend to support the installation of rigorous standards at the early stages of AI development, because failures would be damaging to entire industries, not just to the perpetrator and victim.

The question of *enforcing* policies for beneficial AI is more problematic, given our lack of success in containing malware. In Samuel Butler's *Erewhon* and in Frank Herbert's *Dune*, the solution is to ban all intelligent machines, as a matter of both law and cultural imperative. Perhaps if we find institutional solutions to the malware problem, we will be able to devise some less drastic approach for AI.

References

Aghion, P., Jones, B. F., and Jones, C. I. (2017), 'Artificial Intelligence and Economic Growth', National Bureau of Economic Research Working Paper 23928.

Agrawal, A., Gans, J., and Goldfarb, A. (eds) (2019), *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research.

Aristotle, *Nicomachean Ethics*, Book III, 3, 1112b.
Arnauld, A. (1662), *La logique, ou l'art de penser*, Paris, Chez Charles Savreux.
Bartley, J. (2017), 'The Greens Endorse a Universal Basic Income. Others Need to Follow', *The Guardian*, 2 June.
Bellman, R. E. (1952), 'On the Theory of Dynamic Programming', *Proceedings of the National Academy of Sciences*, **38**(8), 716–19.
— (1957), *Dynamic Programming*, Princeton, NJ, Princeton University Press.
Bernoulli, D. (1738), 'Specimen theoriae novae de mensura sortis', *Proceedings of the St Petersburg Imperial Academy of Sciences*, **5**, 175–92.
Bessen, J. (2019), 'Artificial Intelligence and Jobs: The Role of Demand', in A. Agrawal, J. Gans, and A. Goldfarb (eds), *The Economics of Artificial Intelligence: An Agenda*, National Bureau of Economic Research.
Bowman, S. (2013), 'The Ideal Welfare System is a Basic Income', Welfare and Pensions blog, 25 November, Adam Smith Institute.
Brynjolfsson, E., and McAfee, A. (2016), *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York, W. W. Norton.
Carter, B., Jain, S., Mueller, J., and Gifford, D. (2020), 'Overinterpretation Reveals Image Classification Model Pathologies', arXiv:2003.08907.
Chace, C. (2016), *The Economic Singularity: Artificial Intelligence and the Death of Capitalism*, Three Cs Press.
Chan, L., Hadfield-Menell, D., Srinivasa, S., and Dragan, A. (2019), 'The Assistive Multi-armed Bandit', in Proceedings of Fourteenth ACM/IEEE International Conference on Human–Robot Interaction.
D'Amour, A., Heller, K., Moldovan, D., et al. (2020), 'Underspecification Presents Challenges for Credibility in Modern Machine Learning', arXiv:2011.03395
De Montmort, P. R. (1713), *Essay d'analyse sur les jeux de hazard*, 2nd edn, Paris, Chez Jacques Quillau.
Fickinger, A., Hadfield-Menell, D., Critch, A., and Russell, S. (2020), 'Multi-Principal Assistance Games: Definition and Collegial Mechanisms', in Proceedings of NeurIPS Workshop on Cooperative AI.
Ford, M. (2015), *Rise of the Robots: Technology and the Threat of a Jobless Future*, New York, Basic Books.
Gleave, A., Dennis, M., Kant, N., Wild, C., Levine, S., and Russell, S. (2020), 'Adversarial Policies: Attacking Deep Reinforcement Learning', in Proceedings of Eighth International Conference on Learning Representations (ICLR-20).
Goodman, N. D., Mansinghka, V. K., Roy, D., Bonawitz, K., and Tenenbaum, J. B. (2008), 'Church: A Language for Generative Models', in Proceedings of Twenty-Fourth Conference on Uncertainty in Artificial Intelligence (UAI-08).
Grace, K., Salvatier, J., Dafoe, A., Zhang, B., and Evans, O. (2018), 'When will AI Exceed Human Performance? Evidence from AI Experts', arXiv:1705.08807.
Groth, O., Nitzberg, M., and Russell, S. (2019), 'AI Algorithms Need FDA-style Drug Trials', *Wired*, 15 August.
Hadfield-Menell, D., Dragan, A. D., Abbeel, P., and Russell, S. (2017*a*), 'Cooperative Inverse Reinforcement Learning', in *Advances in Neural Information Processing Systems*, **29**.
— — — — (2017*b*), 'The Off-switch Game', in Proceedings of Twenty-Sixth International Joint Conference on Artificial Intelligence.
Harsanyi, J. (1977), 'Morality and the Theory of Rational Behavior', *Social Research*, **44**, 623–56.
Hillis, D. (2019), 'The First Machine Intelligences', in John Brockman (ed.), *Possible Minds: Twenty-Five Ways of Looking at AI*, New York, Penguin Press.
Hirsch, F. (1977), *The Social Limits to Growth*, Abingdon, Routledge & Kegan Paul.
Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. (2018), 'Real-time Bidding with Multi-agent Reinforcement Learning in Display Advertising', arXiv:1802.09756v2.
Keynes, J. M. (1930), 'Economic Possibilities for our Grandchildren', reprinted in *Essays in Persuasion* (1932), New York, Harcourt Brace.
Koller, D., McAllester, D. A., and Pfeffer, A. (1997), 'Effective Bayesian Inference for Stochastic Programs', in Proceedings of Fourteenth National Conference on Artificial Intelligence (AAAI-97).

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2013), 'ImageNet Classification with Deep Convolutional Neural Networks', in *Advances in Neural Information Processing Systems*, **25**.

Le Bras, R., Arora, N., Kushida, N., Mialle, P., Bondar, I., Tomuta, E., Alamneh, F. K., Feitio, P., Villarroel, M., Vera, B., Sudakov, A., Laban, S., Nippress, S., Bowers, D., Russell, S., and Taylor, T. (2020), 'NET-VISA from Cradle to Adulthood. A Machine-learning Tool for Seismo-acoustic Automatic Association', *Pure and Applied Geophysics*, doi 10.1007/s00024-020-02508-x.

LeCun, Y., Bengio, Y., and Hinton, G. E. (2015), 'Deep Learning', *Nature*, **521**, 436–44.

— Jackel, L., Boser, B., and Denker, J. (1989), 'Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning', *IEEE Communications Magazine*, **27**, 41–6.

Milch, B., Marthi, B., Sontag, D., Russell, S. J., Ong, D., and Kolobov, A. (2005), 'BLOG: Probabilistic Models with Unknown Objects', in Proceedings of Nineteenth International Joint Conference on Artificial Intelligence (IJCAI-05).

Ng, A. Y., and Russell, S. (2000), 'Algorithms for Inverse Reinforcement Learning', in Proceedings of Seventeenth International Conference on Machine Learning.

Nozick, R. (1974), *Anarchy, State, and Utopia*, New York, Basic Books.

Pearl, J. (1988), *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Burlington, MA, Morgan Kaufmann.

— (2000), *Causality: Models, Reasoning, and Inference*, Cambridge, Cambridge University Press.

— McKenzie, D. (2018), *The Book of Why*, New York, Basic Books.

Pettigrew, R. (2020), *Choosing for Changing Selves*, Oxford, Oxford University Press.

Pfeffer, A. (2001), 'IBAL: A Probabilistic Rational Programming Language', in Proceedings of Seventeenth International Joint Conference on Artificial Intelligence (IJCAI-01).

Quillen, D., Jang, E., Nachum, O., Finn, C., Ibarz, J., and Levine, S. (2018), 'Deep Reinforcement Learning for Vision-based Robotic Grasping: A Simulated Comparative Evaluation of Off-policy Methods', in Proceedings of Thirty-Fifth IEEE Conference on Robotics and Automation (ICRA-18).

Russell, S. (1998), 'Learning Agents for Uncertain Environments', in Proceedings of Eleventh ACM Conference on Computational Learning Theory.

— (2019), *Human Compatible: AI and the Problem of Control*, London, Penguin.

— Norvig, P. (2020), *Artificial Intelligence: A Modern Approach*, 4th edn, New York, Pearson.

Samuel, A. (1959), 'Some Studies in Machine Learning Using the Game of Checkers', *IBM Journal of Research and Development*, **3**, 210–29.

Sen, A. (1999), 'The Possibility of Social Choice', *American Economic Review*, **89**, 349–78.

Shah, R., Krasheninnikov, D., Alexander, J., Abbeel, P., and Dragan, A. (2019), 'The Implicit Preference Information in an Initial State', in *Proceedings of Seventh International Conference on Learning Representations*.

Shapley, S. (1953), 'Stochastic Games', *Proceedings of the National Academy of Sciences*, **39**, 1095–100.

Shenker, I. (1977), 'Brainy Robots in our Future, Experts Think', *Detroit Free Press*, 30 September.

Silver, D., Huang, A., Maddison, C. J., Guez, A., and Hassabis, D. (2016), 'Mastering the Game of Go with Deep Neural Networks and Tree Search', *Nature*, **529**, 484–9.

Sullivan, D. (2020), 'A Reintroduction to our Knowledge Graph and Knowledge Panels', Google blog post, 20 May, https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/

Turing, A. (1950), 'Computing Machinery and Intelligence', *Mind*, **59**, 433–60.

Van Zanden, J. L. (ed.) (2014), *How Was Life? Global Well-Being Since 1820*, Paris, OECD Publishing.

Veblen, T. (1899), *The Theory of the Leisure Class: An Economic Study of Institutions*, London, Macmillan.

Von Neumann, J., and Morgenstern, O. (1944), *Theory of Games and Economic Behavior*, Princeton, NJ, Princeton University Press.