

The MIT Press

If We Succeed

Author(s): Stuart Russell

Source: *Daedalus*, Spring 2022, Vol. 151, No. 2, AI & Society (Spring 2022), pp. 43-57

Published by: The MIT Press on behalf of American Academy of Arts & Sciences

Stable URL: https://www.jstor.org/stable/10.2307/48662025

JSTOR is a not-for-profit service that helps scholars, researchers, and students discover, use, and build upon a wide range of content in a trusted digital archive. We use information technology and tools to increase productivity and facilitate new forms of scholarship. For more information about JSTOR, please contact support@jstor.org.

Your use of the JSTOR archive indicates your acceptance of the Terms & Conditions of Use, available at https://about.jstor.org/terms



American Academy of Arts & Sciences and *The MIT Press* are collaborating with JSTOR to digitize, preserve and extend access to *Daedalus*

This content downloaded from 12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

If We Succeed

Stuart Russell

Since its inception, AI has operated within a standard model whereby systems are designed to optimize a fixed, known objective. This model has 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 society. At the same time, the standard model will become progressively 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 behavior that are more robust, controllable, and deferential.

The central technical concept in AI is that of an agent: an entity that perceives and acts.[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 that mean?

In answering this question, AI has drawn on a much longer train of thought concerning rational behavior: what is the right thing to do? Aristotle 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."[2] 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. In the seventeenth century, theologian and philosopher Antoine Arnauld broadened Aristotle's theory to include uncertainty in a quantitative way, proposing that we should act to maximize the expected value of the outcome (that is, averaging the values of different possible outcomes weighted by their probabilities).[3] In the eighteenth century, Swiss mathematician Daniel Bernoulli refined the notion of value, moving it from an external quantity (typically money) to an internal quantity that he called utili-

© 2022 by Stuart Russell
Published under a Creative Commons Attribution-NonCommercial 4.0 International (CC BY-NC 4.0) license
https://doi.org/10.1162/DAED_a_01899

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

ty.[4] French mathematician Pierre Rémond de Montmort noted that in games (decision situations involving two or more agents) a rational agent might have to act randomly to avoid being second-guessed.[5] And in the twentieth century, mathematician John Von Neumann and economist Oskar Morgenstern tied all these ideas together into an axiomatic framework: rational agents must satisfy certain properties such as transitivity of preferences (if you prefer A to B and B to C, you must prefer A to C), and any agent satisfying those properties can be viewed as having a utility function on states and choosing actions that maximize expected utility.[6]

As AI emerged alongside computer science in the 1940s and 1950s, researchers 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 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 maximize the utility of individuals or the welfare of groups.

Within the standard model, new ideas have arisen fairly regularly since the 1950s, leading eventually to impressive real-world applications. 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 such as the Deep Blue chess program, which defeated world champion Garry Kasparov in 1997, and AlphaGo, which defeated world Go champion Ke Jie in 2017. 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 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 automated reasoning has advanced dramatically. For example, it is now routine to verify the correctness of VLSI (very large scale integration) chip designs before production and the correctness of software systems and cybersecurity

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

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 five hundred billion facts about five billion entities and is used to answer directly more than one-third of all queries submitted to the Google search engine.[7]

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 computer scientist Judea Pearl and others.[8] 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 modeling, diagnosis, monitoring, and prediction of a wide range of complex systems, including jet engines, Mars rovers, ecological networks, and intensive care protocols. Causal networks, 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.[9]

The development of probabilistic programming languages, or PPLs, provides a universal representation for probability models, meaning that any model representable in any formalism can be represented efficiently in a PPL.[10] 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; while operating, it may dynamically construct internal representations involving hundreds of thousands of random variables.[11]

Alan Turing suggested that machine learning would be the most practical way to create AI capabilities.[12] The most common paradigm – one shared with statistical prediction methods – is supervised learning, wherein labeled examples are provided to a learning algorithm that outputs a predictive hypothesis with which to la-

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

bel unlabeled 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.[13] 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 (Generative Pre-trained Transformer 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.[14] 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 and operations research.[15] Algorithms developed in these fields typically handle only small problems with up to one million states. In AI, the development of reinforcement learning (RL) has allowed researchers to address much larger problems satisfactorily, including backgammon with 10^{19} positions and Go with 10^{170} positions.[16] 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 to improving the ability of robots to grasp previously unseen objects.[17] As with supervised learning, applications of deep networks in RL may also be quite fragile.[18]

With modest advances in perception and dexterity, we can expect to see robots moving into a variety of unstructured environments, including roads, warehouses, agriculture, mining, and warfare. We may see progress on language understanding comparable to the progress on image understanding made over 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, in a shallow sense, understanding almost everything the human race has ever written. And text would be augmented by satellite imagery, enabling computers to see every object (fifty centimeters or larger) on Earth every day, weather permitting.

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 toward 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 per-

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

ceptual data. (This approach underlies, for example, Waymo's industry-leading self-driving car project.) The reasons for this prediction are complex, but include 1) the performance problems with deep learning mentioned earlier; 2) the possibility that such problems may contribute to the failure of flagship projects such as self-driving cars; 3) the advantages, in terms of rigor, transparency, and modularity, of being able to analyze systems as possessing knowledge and reasoning with that knowledge; 4) the expressive limitations of circuit-based representations (including deep learning systems) for capturing general knowledge; 5) the essential role played by prior knowledge in enabling a learning system to generalize robustly from small numbers of examples; and 6) 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.

Even in its present state, the technology of artificial intelligence raises many concerns as it transitions from research into widespread use. These concerns include potential misuses such as cybercrime, surveillance, disinformation, and political manipulation; the exacerbation of inequality and of many forms of bias in society; the creation and deployment of lethal autonomous weapons; and the usurpation of human roles in the economy and in social relationships.

These issues are addressed admirably in the other essays in this volume, many of which contribute to an important yet lamentably only recent trend: understanding potential applications of AI not only as technological problems to be solved, but also as existing in a social context. Success is to be measured not by the accuracy of the AI system's predictions and decisions, but by the real-world consequences of deploying the system. In other words, we need a theory of sociotechnical embedding for AI systems, somewhat analogous to the role that city planning plays for the artifacts produced by civil engineering and architecture. Absent such a theory, we are left with the market to sort through different systems and embeddings. For all sorts of reasons, including network effects and social externalities, this is unlikely to work.[19]

My concern here, however, is with the potential consequences of success in creating general-purpose AI: that is, systems capable of quickly learning to perform at a high level in any task environment where humans (or collections of humans) can perform well. General-purpose AI has been the long-term goal of the field since its inception. For example, Herbert Simon and Allen Newell, two pioneers of AI research, famously predicted in 1957: "There are now in the world machines that think, that learn and that create. Moreover, their

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

ability to do these things is going to increase rapidly until – in a visible future – the range of problems they can handle will be coextensive with the range to which the human mind has been applied."[20]

It would be an oversimplification to view progress in AI as occurring along a one-dimensional, numerical scale of "intelligence." While such a scale has some relevance for humans, AI capabilities in different branches of cognitive activity vary so markedly as to make a single scale completely inapplicable. For example, a search engine remembers very well and cannot plan at all; a chess program plans very well and cannot remember at all. For this reason, there will be no single moment at which AI "exceeds human intelligence." By the time that AI systems exhibit generality across all branches, direct comparisons to humans will be meaningless. Almost certainly, such systems would already far exceed human capabilities in many areas thanks to the massive speed, memory, and input bandwidth advantages of computers compared with humans.

That is not to imply that we are close to achieving general-purpose AI. Suggestions that we simply need to collect more data or acquire more computing power seem overly optimistic. For example, current natural-language systems process, in only a few days, thousands of times more text than any human has ever read, yet their understanding of language is brittle and often parrot-like. We need conceptual breakthroughs in a number of areas besides language understanding, 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 earliest pioneers in 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."[21] 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.[22]

Given the huge levels of investment in AI research and development and the influx of talented researchers into the field, it is reasonable to suppose that fundamental advances will continue to occur as we find new applications for which existing techniques and concepts are inadequate. As noted above, these advances are hard to predict, but there are no fundamental obstacles that prevent them from occurring. Indeed, what evidence could there be that no physically possible arrangement of atoms can outperform the human brain?

The potential benefits of general-purpose AI would be far greater than those of a collection of narrow, application-specific AI systems. For this reason, the prospect of creating general-purpose AI is driving massive investments and geopolitical rivalries.

One can speculate about solving major open problems, such as extending human life indefinitely or developing faster-than-light travel, but these staples of sci-

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

ence 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 considered respectable in a developed country. Choosing "respectable" (somewhat arbitrarily) to mean the eighty-eighth percentile in the United States, this goal represents an almost tenfold increase in global GDP, from $76 trillion to $750 trillion per year. The increased income stream resulting from this achievement has a net present value of $13.5 quadrillion, assuming a discount factor of 5 percent. (The value is $9.4 quadrillion or $6.8 quadrillion if the technology is phased in over ten or twenty years.) These numbers tower over the amounts currently invested in AI research, and momentum toward this goal will increase as technical advances bring general-purpose AI closer to realization.

Such a tenfold increase in global GDP per capita took place over 190 years, from 1820 to 2010.[23] It 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, 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 – 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 one hundred guests, running elections, teaching children to read, or doing whatever else needs doing. It is the generality of general-purpose intelligence that makes this possible.

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.[24] One can also expect finite resources such as land, human attention, and perhaps raw materials to become relatively more expensive.

The incentives for further development of AI, then, are huge, and the momentum appears unstoppable. We must, therefore, ask, "What if we succeed?" This question is seldom considered in the AI literature, which is focused primarily on the pursuit of success rather than on its consequences. Alan

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

Turing, widely regarded as the founder of computer science, did consider the question. And in 1951, during a lecture given to a learned society in Manchester, he answered: "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."[25]

Turing's prediction is a natural response to the following conundrum: our intelligence gives us power over the world and over other species; we will build systems with superhuman intelligence; therefore, we face the problem of retaining power, forever, over entities that are far more powerful than ourselves.

Within the standard model of AI, the meaning of "power" is clear: the ability to achieve one's objectives regardless of the objectives and actions of others. I believe the future Turing had in mind was one in which machines take control as a result of pursing fixed objectives that are misaligned with human benefit. These fixed objectives will be ones that we ourselves have inserted: there is no need to posit some form of emergent consciousness that spontaneously generates its own objectives. All that is needed to assure catastrophe is a highly competent machine combined with humans who have an imperfect ability to specify human preferences completely and correctly. This is why, when a genie has granted us three wishes, our third wish is always to undo the first two wishes.

Unfortunately, the standard model within which almost all current AI systems are developed makes this future almost inevitable. 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 machines take preemptive steps to ensure that the stated objective is achieved.

The standard model, then, despite all its achievements, is a mistake. The mistake comes from transferring a perfectly reasonable definition of intelligence from humans to machines. It is not rational for humans to deploy machines that pursue fixed objectives when there is a significant possibility that those objectives diverge from our own.

A more sensible definition of AI would have machines pursuing *our* objectives. Of course, our objectives – in more technical language, our preferences among lotteries over complete futures – are in us, and not in the machines. This means that machines will necessarily be *uncertain* about our objectives, while being obliged to pursue them on our behalf. In this pursuit, they will be aided by evidence concerning human preferences. This evidence comes from human behavior, broadly construed, including choices, inaction, commands, requests, guidance, permissions, artifacts, and social structures.

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

This new model for AI, with its emphasis on uncertainty about objectives, entails a binary coupling between machines and humans that gives it a flavor quite different from the unary standard model of decoupled machines pursuing fixed objectives. The standard model can be viewed as an extreme special case of the new model, applicable only when it is reasonable to suppose that, within the machine's scope of action, the relevant human objectives can be specified completely and correctly. It turns out that the uncertainty inherent in the new model is crucial to building AI systems of arbitrary intelligence that are provably beneficial to humans.

Uncertainty concerning objectives is a surprisingly understudied topic. In the 1980s, the AI community acknowledged the inevitability of uncertainty concerning the current state and the effects of actions, but we continued to assume perfect knowledge of the objective. For artificially defined puzzles and games, this may be appropriate, but for other problems, such as recommending medical treatments, it is clear that the relevant preferences (of patients, families, doctors, insurers, hospital systems, taxpayers, and so on) 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 additional evidence of the true objectives can be acquired. 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 basic idea is made more precise in the framework of assistance games, originally known as cooperative inverse reinforcement learning (CIRL) games.[26] 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, humans who do not know their own preferences, multiple human participants, and multiple robots, among other variations.[27] Human actions in such games can, of course, include communicative actions such as stating preferences, making requests, and issuing commands.

Assistance games are connected to inverse reinforcement learning (IRL) because the robot can learn more about human preferences from the observation of human behavior – a process that is the dual of reinforcement learning, wherein behavior is learned from rewards and punishments.[28] 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 their preferences, and the robot may interpret the human's actions in this light, rather than simply as demonstrations of optimal behavior.

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

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 probability distribution over human rewards, one can show that a robot solving an assistance game has nonnegative value to humans.[29]
- A robot that is uncertain about the human's preferences has a nonnegative incentive to allow itself to be switched off.[30] 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" behavior to benefit the human.[31] 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.

Needless to say, there are many open research problems in the new model of AI. First, we need to examine each existing research area (search, game playing, constraint satisfaction, planning, reinforcement learning, and so on) and remove the assumption of a fixed, known objective, rebuilding that area on a broader foundation that allows for uncertainty about objectives. The key questions in each area are how to formulate the machine's initial uncertainty about human preferences and how to codify the run-time flow of preference information from human to machine.

Another set of research problems arises when we consider how the machine can learn about human preferences from human behavior in the assistance game. The first difficulty is that humans are irrational in the sense that our actions do not reflect our preferences. This irrationality arises in part from our computational limitations relative to the complexity of the decisions we face. For example, if two humans are playing chess and one of them loses, it is because the loser (and possibly the winner, too) made a mistake, a move that led inevitably to a forced loss. A machine observing that move and assuming perfect rationality on the part of the human might well conclude that the human *preferred* to lose. Thus, to avoid reaching such conclusions, the machine must take into account the *actual* cognitive mechanisms of humans.

Another important consequence of human computational limitations is that they force us to organize our behavior hierarchically. That is, we make (defeasible) commitments to higher-level goals such as "write an essay on a human-compatible approach to AI." Then, rather than considering all possible sequences of words, from "aardvark aardvark aardvark" to "zyzzyva zyzzyva zyzzyva," as a chess program might do, we choose among subtasks such as "write the introduc-

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

tion" and "read more about preference elicitation." Eventually, we get down to the choice of words, and then typing each word involves a sequence of keystrokes, each of which is in turn a sequence of motor control commands to the muscles of the arms and hands. At any given point, then, a human is embedded at various particular levels of multiple deep and complex hierarchies of partially overlapping activities and subgoals. This means that for the machine to understand human actions, it probably needs to understand a good deal about what these hierarchies are and how we use them to navigate the real world.

Other research problems engage directly with philosophy and the social sciences. For example, there is the question of social aggregation, a staple of economics and moral philosophy: how should a machine make decisions when its actions affect the interests of more than one human being? Issues include the preferences of evil individuals, relative preferences and positional goods, and interpersonal comparison of preferences.[32]

Also of great importance is the plasticity of human preferences: the fact that they seem to change over time as the result of experiences. It is hard to explain how such changes can be made rationally, since they make one's future self less likely to satisfy one's present preferences about the future. Yet plasticity seems fundamentally important to the entire enterprise, because newborn infants certainly lack the rich, nuanced, culturally informed preference structures of adults. Indeed, it seems likely that our preferences are at least partially formed by a process resembling inverse reinforcement learning, whereby we absorb preferences that explain the behavior of those around us. Such a process would tend to give cultures some degree of autonomy from the otherwise homogenizing effects of our dopamine-based reward system.

Plasticity also raises the obvious question of which human H the machine should try to help: H_{2022}, H_{2035}, or some time-averaged H?[33] Plasticity is also problematic because of the possibility that the machine may, by subtly influencing the environment, gradually mold H's preferences in directions that make them easier to satisfy. This problem is a familiar one in human society, where culture and propaganda mold the preferences of humans to facilitate their compliance with existing power structures.

Let us assume, for the sake of argument, that all these obstacles can be overcome, as well as all of the obstacles to the development of truly capable AI systems. Are we then home free? Would provably beneficial, superintelligent AI usher in a golden age for humanity? Not necessarily. There remains the issue of adoption: how can we obtain broad agreement on suitable design principles, and how can we ensure that only suitably designed AI systems are deployed?

On the question of obtaining agreement at the policy level, it is necessary first to generate consensus within the research community on the basic ideas of – and

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

design templates for – provably beneficial AI, so that policy-makers have some concrete guidance on what sorts of regulations might make sense. Economic incentives would tend to support the installation of rigorous standards at the early stages of AI development, since failures would be damaging to entire industries, not just to the perpetrator and victim. We already see this in miniature with the imposition of machine-checkable software standards for cell phone applications.

On the question of enforcement, I am less sanguine. If the next Dr. Evil wants to take over the world, he or she might remove the safety catch, so to speak, and deploy a poorly designed AI system that ends up destroying the world instead. This is a hugely magnified version of the problem we currently face with malware. Our track record in solving the latter problem does not provide grounds for optimism concerning the former. 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 regulating AI.

The problem of misuse is not limited to evil masterminds. One possible future for humanity in the age of superintelligent AI is that of a race of lotus eaters, progressively enfeebled as machines take over the management of our entire civilization. This is the future imagined in E. M. Forster's story *The Machine Stops*, written in 1909. We may say, now, that such a future is undesirable; the machines may agree with us and volunteer to stand back, requiring humanity to exert itself and maintain its vigor. But exertion is tiring, and we may, in our usual myopic way, design AI systems that are not quite so concerned about the long-term vigor of humanity and are just a little more helpful than they would otherwise wish to be. Unfortunately, this slope is very slippery indeed.

Finding a solution to the AI control problem is an important task; it may be, in the words of philosopher Nick Bostrom, "the essential task of our age."[34] Up to now, AI research has focused on systems that are better at making decisions, but this is not the same as making better decisions if human and machine objectives diverge.

This problem requires a change in the definition of AI itself: from a field concerned with a unary notion of intelligence as the optimization of a given objective to a field concerned with a binary notion of machines that are provably beneficial for humans. Taking the problem seriously seems likely to yield new ways of thinking about AI, its purpose, and our relationship with it.

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

ABOUT THE AUTHOR

Stuart Russell is Professor of Computer Science and the Smith-Zadeh Professor in Engineering at the University of California, Berkeley, and Honorary Fellow at Wadham College, Oxford. He is the author of *Artificial Intelligence: A Modern Approach* (with Peter Norvig, 4th ed., 2021), *Human Compatible: AI and the Problem of Control* (2019), and *Do the Right Thing: Studies in Limited Rationality* (with Eric H. Wefald, 1991).

ENDNOTES

[1] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 4th ed. (Hoboken, N.J.: Pearson, 2021).

[2] Aristotle, *Nicomachean Ethics* 3.3, 1112b.

[3] Antoine Arnauld, *La logique, ou l'art de penser* (Paris: Chez Charles Savreux, 1662).

[4] Daniel Bernoulli, "Specimen theoriae novae de mensura sortis," *Proceedings of the St. Petersburg Imperial Academy of Sciences* 5 (1738): 175–192.

[5] Pierre Rémond de Montmort, *Essay d'analyse sur les jeux de hazard*, 2nd ed. (Paris: Chez Jacques Quillau, 1713).

[6] John Von Neumann and Oskar Morgenstern, *Theory of Games and Economic Behavior* (Princeton, N.J.: Princeton University Press, 1944).

[7] Danny Sullivan "A Reintroduction to Our Knowledge Graph and Knowledge Panels," The Keyword, Google, May 20, 2020, https://blog.google/products/search/about-knowledge-graph-and-knowledge-panels/.

[8] Judea Pearl, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference* (Burlington, Mass.: Morgan Kaufmann, 1988).

[9] Judea Pearl, *Causality: Models, Reasoning, and Inference* (Cambridge: Cambridge University Press, 2000); and Judea Pearl and Dana Mackenzie, *The Book of Why: The New Science of Cause and Effect* (New York: Basic Books, 2018).

[10] Daphne Koller, David McAllester, and Avi Pfeffer, "Effective Bayesian Inference for Stochastic Programs," in *Proceedings of Fourteenth National Conference on Artificial Intelligence* (Menlo Park, Calif.: Association for the Advancement of Artificial Intelligence, 1997); Avi Pfeffer, "IBAL: A Probabilistic Rational Programming Language," in *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2001); Brian Milch, Bhaskara Marthi, Stuart Russell, et al., "BLOG: Probabilistic Models with Unknown Objects," in *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2005); and Noah D. Goodman, Vikash K. Mansinghka, Daniel Roy, et al., "Church: A Language for Generative Models," in *Proceedings of the Twenty-Fourth Conference on Uncertainty in Artificial Intelligence* (Helsinki: Association for Uncertainty in Artificial Intelligence, 2008).

[11] Ronan Le Bras, Nimar Arora, Noriyuki Kushida, et al., "NET-VISA from Cradle to Adulthood: A Machine-Learning Tool for Seismo-Acoustic Automatic Association," *Pure and Applied Geophysics* 178 (2021): 2437–2458.

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

[12] Alan Turing, "Computing Machinery and Intelligence," *Mind* 56 (236) (1950): 43–60.

[13] Yann LeCun, Lawrence Jackel, Bernhard Boser, et al., "Handwritten Digit Recognition: Applications of Neural Network Chips and Automatic Learning," *IEEE Communications Magazine* 27 (11) (1989): 41–46; Yann LeCun, Yoshua Bengio, and Geoffrey Hinton, "Deep Learning," *Nature* 521 (7553) (2015): 436–444; Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Advances in Neural Information Processing Systems* 25 (2) (2012): 1097–1105.

[14] Brandon Carter, Siddhartha Jain, Jonas Mueller, and David Gifford, "Overinterpretation Reveals Image Classification Model Pathologies," arXiv (2020), https://arxiv.org/abs/2003.08907; and Alexander D'Amour, Katherine Heller, Dan Moldovan, et al., "Underspecification Presents Challenges for Credibility in Modern Machine Learning," arXiv (2020), https://arxiv.org/abs/2011.03395.

[15] Lloyd S. Shapley, "Stochastic Games," *Proceedings of the National Academy of Sciences* 39 (10) (1953): 1095–1100; Richard Bellman, "On the Theory of Dynamic Programming," *Proceedings of the National Academy of Sciences* 38 (8) (1952): 716–719; and Richard Bellman, *Dynamic Programming* (Princeton, N.J.: Princeton University Press, 1957).

[16] Arthur L. Samuel, "Some Studies in Machine Learning Using the Game of Checkers," *IBM Journal of Research and Development* 3 (3) (1959): 210–229; and David Silver, Aja Huang, Chris J. Maddison, et al., "Mastering the Game of Go with Deep Neural Networks and Tree Search," *Nature* 529 (7587) (2016): 484–489.

[17] Junqi Jin, Chengru Song, Han Li, et al., "Real-Time Bidding with Multi-Agent Reinforcement Learning in Display Advertising," in *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (New York: Association for Computing Machinery, 2018), 2193–2201; and Deirdre Quillen, Eric Jang, Ofir Nachum, et al., "Deep Reinforcement Learning for Vision-Based Robotic Grasping: A Simulated Comparative Evaluation of Off-Policy Methods," in *Proceedings of the 2018 IEEE International Conference on Robotics and Automation (ICRA)* (Piscataway, N.J.: Institute of Electrical and Electronics Engineers, 2018), 6284–6291.

[18] Adam Gleave, Michael Dennis, Neel Kant, et al., "Adversarial Policies: Attacking Deep Reinforcement Learning," in *Proceedings of the Eighth International Conference on Learning Representations* (La Jolla, Calif.: International Conference on Learning Representations, 2020).

[19] Eric Posner and Glen Weyl, *Radical Markets: Uprooting Capitalism and Democracy for a Just Society* (Princeton, N.J.: Princeton University Press, 2019).

[20] Herbert A. Simon and Allen Newell, "Heuristic Problem Solving: The Next Advance in Operations Research," *Operations Research* 6 (1) (1958).

[21] Israel Shenker, "Brainy Robots in Our Future, Experts Think," *Detroit Free Press*, September 30, 1977.

[22] Katja Grace, John Salvatier, Allan Dafoe, et al., "When Will AI Exceed Human Performance? Evidence from AI Experts," *Journal of Artificial Intelligence Research* (62) (2018): 729–754.

[23] Jan Luiten Van Zanden, Joerg Baten, Marco Mira d'Ercole, et al., eds., *How Was Life? Global Well-Being Since 1820* (Paris: OECD Publishing, 2014).

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms

[24] Philippe Aghion, Benjamin F. Jones, and Charles I. Jones, "Artificial Intelligence and Economic Growth," National Bureau of Economic Research Working Paper 23928 (Cambridge, Mass.: National Bureau of Economic Research, 2017).

[25] Alan Turing, "'Intelligent Machinery, A Heretical Theory,' a Lecture Given to '51 Society' at Manchester," AMT/B/4, The Turing Digital Archive, https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-4.

[26] Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "Cooperative Inverse Reinforcement Learning," *Advances in Neural Information Processing Systems* 29 (2016): 3909–3917.

[27] For examples, see Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel, and Stuart Russell, "The Off-Switch Game," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (Santa Clara, Calif.: International Joint Conference on Artificial Intelligence Organization, 2017), 220–227; Lawrence Chan, Dylan Hadfield-Menell, Siddhartha Srinivasa, and Anca Dragan, "The Assistive Multi-Armed Bandit," in *14th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2019)* (Red Hook, N.Y.: Curran Associates, Inc., 2019), 354–363; and Arnaud Fickinger, Simon Zhuang, Andrew Critch, et al., "Multi-Principal Assistance Games: Definition and Collegial Mechanisms," presented at the Cooperative AI Research Workshop at the 34th Conference on Neural Information Processing Systems (NeurIPS 2020), virtual conference, December 6–12, 2020.

[28] Stuart Russell, "Learning Agents for Uncertain Environments," in *Proceedings of the Eleventh ACM Conference on Computational Learning Theory* (New York: Association for Computing Machinery, 1998); and Andrew Ng and Stuart Russell, "Algorithms for Inverse Reinforcement Learning," in *Proceedings of Seventeenth International Conference on Machine Learning* (San Francisco: Morgan Kaufmann Publishers, Inc., 2000).

[29] Hadfield-Menell et al., "Cooperative Inverse Reinforcement Learning."

[30] Hadfield-Menell et al., "The Off-Switch Game."

[31] Rohin Shah, Dmitrii Krasheninnikov, Jordan Alexander, et al., "Preferences Implicit in the State of the World," in *Proceedings of the Seventh International Conference on Learning Representations* (La Jolla, Calif.: International Conference on Learning Representations, 2019).

[32] John C. Harsanyi, "Morality and the Theory of Rational Behavior," *Social Research* 44 (4) (1977): 623–656; Thorstein Veblen, *The Theory of the Leisure Class: An Economic Study of Institutions* (London: Macmillan Company, 1899); Fred Hirsch, *Social Limits of Growth* (London: Routledge and Kegan Paul, 1977); Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974); and Amartya Sen, "The Possibility of Social Choice," *American Economic Review* 89 (3) (1999): 349–378.

[33] Richard Pettigrew, *Choosing for Changing Selves* (Oxford: Oxford University Press, 2020).

[34] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford: Oxford University Press, 2014).

This content downloaded from
12.13.248.126 on Fri, 08 Sep 2023 16:21:28 +00:00
All use subject to https://about.jstor.org/terms