Artificial Intelligence and the Problem of Control

Stuart Russell

Abstract A long tradition in philosophy and economics equates intelligence with the ability to act rationally—that is, to choose actions that can be expected to achieve one’s objectives. This framework 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. Unfortunately, the standard model is unworkable as a foundation for further progress because it is seldom possible to specify objectives completely and correctly in the real world. The chapter proposes 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 to humans.

1 The Standard Model

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 (*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,

¹The word “agent” in AI carries no connotation of acting on behalf of another.

S. Russell (✉)
University of California, Berkeley, CA, USA
e-mail: russell@berkeley.edu

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 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.

In short, the standard model of AI (and related disciplines) is a pillar of twentieth-century technology.

2 Difficulties of the Standard Model

Unfortunately, the standard model is unworkable as a foundation for further progress. Once AI systems move out of the laboratory (or artificially defined environments such as the simulated chessboard) 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 (1951): “*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.*” We can expect a sufficiently capable machine pursuing a fixed objective to take preemptive steps to ensure that the stated objective is achieved, including acquiring physical and

computational resource and defending against any possible attempt to interfere with goal achievement.

The Vienna Manifesto on Digital Humanism includes the following principle: “We must shape technologies in accordance with human values and needs, instead of allowing technologies to shape humans.” Perhaps the clearest example demonstrating the need for this principle is given by machine learning algorithms performing content selection on social media platforms. Such algorithms typically pursue the objective of maximizing clickthrough or a related metric. Rather than simply adjusting their recommendations to suit human preferences, these algorithms will, in pursuit of their long-term objective, learn to manipulate humans to make them more predictable in their clicking behavior (Groth et al. 2019).² This effect may be contributing to growing polarization and extremism in many countries.

The mistake in the standard model 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, 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—turns out to be crucial to building AI systems of arbitrary intelligence that are *provably beneficial* to humans. In other words, I propose to do more than “shape technologies in accordance with human values and needs.” Because we cannot necessarily articulate those values and needs, we must design technologies that will, by their very constitution, respond to human values and needs, whatever they are.

3 A New Model

In *Human Compatible* (Russell 2019), I suggest three principles underlying a new model for creating AI systems:

1. The machine’s only objective is to maximize the realization of human preferences.
2. The machine is initially uncertain about what those preferences are.
3. The ultimate source of information about human preferences is human behavior.

²Providing additional evidence for the significance of the problem of misspecified objectives, Hillis (2019) 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.

As noted in the preceding section, the uncertainty about objectives that the second principle espouses is a relatively unstudied concept in AI—yet it is central to ensuring that we not lose control over increasingly capable AI systems.

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 they embraced uncertainty in these aspects of the problem statement. It is not at all clear why, for the most part, they failed to notice that there must also be uncertainty in the objective. Although some AI problems such as puzzle solving are designed to have 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 “runtime.” This flow comes from evidence provided by human behavior, as the third principle asserts.

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. (2017a). 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 (in the basic version) knows the payoff 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 payoff function and updates it as the robot and human interact during the game. The basic assistance game model can be elaborated to allow for imperfectly rational humans (Hadfield-Menell et al. 2017b), 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 behavior—a process that is the dual of reinforcement learning, wherein behavior 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. This two-way process lends the framework an inevitable game-theoretic character that produces, among other phenomena, emergent conventions for communicating preference information.

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 probability distribution over human rewards, one can show that a robot solving an assistance game has non-negative value to humans (Hadfield-Menell et al. 2017a).
- 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. 2017b). 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 (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 most directly relevant to moral philosophy and the social sciences is the question of *social aggregation*: how should a machine decide when its actions affect the interests of more than one human being? 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 policy makers 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. As the Manifesto underscores, the technology of AI has no value in itself beyond its ability to benefit humanity.

References

- Aristotle (n.d.). *Nicomachean Ethics*, Book III, 3, 1112b.
- Arnauld, A. (1662). *La logique, ou l'art de penser*. Paris: Chez Charles Savreux.
- Bernoulli, D. (1738). Specimen theoriae novae de mensura sortis. *Proceedings of the St. Petersburg Imperial Academy of Sciences*, 5, 175–92.
- Chan, L., Hadfield-Menell, D., Srinivasa, S., & Dragan, A. (2019). The assistive multi-armed bandit. In *Proc. Fourteenth ACM/IEEE International Conference on Human–Robot Interaction*.
- De Montmort, P. R. (1713). *Essay d'analyse sur les jeux de hazard*, 2nd ed. Paris: Chez Jacques Quillau.
- Fickinger, A., Hadfield-Menell, D., Critch, A., & Russell, S. (2020). Multi-Principal Assistance Games: Definition and Collegial Mechanisms. In *Proc. NeurIPS Workshop on Cooperative AI*.
- Groth, O., Nitzberg, M., & Russell, S. (2019, August 15). AI algorithms need FDA-style drug trials. *Wired*.
- Hadfield-Menell, D., Dragan, A. D., Abbeel, P., & Russell, S. (2017a). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems* 29.
- Hadfield-Menell, D., Dragan, A. D., Abbeel, P., & Russell, S. (2017b). The off-switch game. In *Proc. Twenty-Sixth International Joint Conference on Artificial Intelligence*.
- Harsanyi, J. (1977). Morality and the theory of rational behavior. *Social Research*, 44, 623–656.
- Hillis, D. (2019). The first machine intelligences. In John Brockman (ed.), *Possible Minds: Twenty-Five Ways of Looking at AI*. Penguin Press.
- Hirsch, F. (1977). *The Social Limits to Growth*. Routledge & Kegan Paul.
- Ng, A. Y. & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proc. Seventeenth International Conference on Machine Learning*.
- Nozick, R. (1974). *Anarchy, State, and Utopia*. Basic Books.
- Pettigrew, R. (2020). *Choosing for Changing Selves*. Oxford University Press.
- Russell, S. (1998). Learning agents for uncertain environments. In *Proc. Eleventh ACM Conference on Computational Learning Theory*.
- Russell, S. (2019). *Human Compatible: AI and the Problem of Control*. London: Penguin.
- Russell, S. & Norvig, P. (2020). *Artificial Intelligence: A Modern Approach* (4th edition). Pearson.
- Sen, A. (1999). The Possibility of Social Choice. *American Economic Review*, 89, 349–378.
- Shah, R., Krashennikov, D., Alexander, J., Abbeel, P., & Dragan, A. (2019). The implicit preference information in an initial state. In *Proc. Seventh International Conference on Learning Representations*.
- Turing, A. (1951). “Can digital machines think?” Radio broadcast, BBC Third Programme. Typescript available at turingarchive.org.
- Veblen, T. (1899). *The Theory of the Leisure Class: An Economic Study of Institutions*. Macmillan.
- von Neumann, J. and Morgenstern, O. (1944). *Theory of Games and Economic Behavior*. Princeton University Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.