Human-Compatible Artificial Intelligence

Stuart Russell
Computer Science Division, University of California, Berkeley

OXFORD
UNIVERSITY PRESS

1

Human-Compatible Artificial Intelligence

1.1 Introduction

Artificial intelligence (AI) has as its aim the creation of intelligent machines. An entity is considered to be intelligent, roughly speaking, if it chooses actions that are expected to achieve its objectives, given what it has perceived.[1] Applying this definition to machines, one can deduce that AI aims to create machines that choose actions that are expected to achieve their objectives, given what they have perceived.

Now, what are these objectives? To be sure, they are——up to now, at least——objectives that we put into them; but, nonetheless, they are objectives that operate exactly as if they were the machines' own and about which they are completely certain. We might call this the *standard model* of AI: build optimizing machines, plug in the objectives, and off they go. This model prevails not just in AI but also in control theory (minimizing a cost function), operations research (maximizing a sum of rewards), economics (maximizing individual utilities, GDP, quarterly profits, or social welfare), and statistics (minimizing a loss function). The standard model is a pillar of twentieth-century technology.

Unfortunately, this standard model is a mistake. It makes no sense to design machines that are beneficial to us *only* if we write down our objectives completely and correctly. If the objective is wrong, we might be lucky and notice the machine's surprisingly objectionable behavior and be able to switch it off in time. Or, if the machine is more intelligent than us, the problem may be irreversible. The more intelligent the machine, the worse the outcome for humans: the machine will have a greater ability to alter the world in ways that are inconsistent with our true objectives and greater skill in foreseeing and preventing any interference with its plans.

In 1960, after seeing Arthur Samuel's checker-playing program learn to play checkers far better than its creator, Norbert Wiener (1960) gave a clear warning:

> If we use, to achieve our purposes, a mechanical agency with whose operation we cannot efficiently interfere ... we had better be quite sure that the purpose put into the machine is

[1] This definition can be elaborated and made more precise in various ways—particularly with respect to whether the choosing and expecting occur within the agent, within the agent's designer, or some combination of both. The latter certainly holds for human agents, viewing evolution as the designer. The word "objective" here is also used informally, and does not refer just to end goals. For most purposes, an adequately general formal definition of "objective" covers preferences over lotteries over complete state sequences. Moreover, "state" here includes mental state as well as the world state external to the entity.

the purpose which we really desire.

Echoes of Wiener's warning can be discerned in contemporary assertions that "superintelligent AI" may present an existential risk to humanity. (In the context of the standard model, "superintelligent" means having a superhuman capacity to achieve given objectives.) Concerns have been raised by such observers as Nick Bostrom (2014), Elon Musk (Kumparak, 2014), Bill Gates (2015),[2] and Stephen Hawking (Osborne, 2017). There is very little chance that as humans 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.

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? The definition of intelligence is *unary*, in the sense that it applies to an entity by itself. 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. Thus, we have a binary definition: entity A chooses actions that are expected to achieve the objectives of entity B, given what entity A has perceived. In the unlikely event that we (entity B) can specify the objectives completely and correctly and insert them into the machine (entity A), then we can recover the original, unary definition. 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. We must, therefore, reconstruct the foundations of AI along binary rather than unary lines.

1.2 Artificial Intelligence

The goal of AI research has been to understand the principles underlying intelligent behavior and to build those principles into machines that can then exhibit such behavior. In the 1960s and 1970s, the prevailing theoretical definition of intelligence was the capacity for logical reasoning, including the ability to derive plans of action guaranteed to achieve a specified goal. A popular variant was the problem-solving paradigm, which requires finding a minimum-cost sequence of actions guaranteed to reach a goal state. More recently, a consensus has emerged in AI around the idea of a rational agent that perceives and acts in order to maximize its expected utility. (In Markov decision processes and reinforcement learning, utility is further decomposed into a sum of rewards accrued through the sequence of transitions in the environment state.) Subfields such as logical planning, robotics, and natural-language understanding are special cases of the general paradigm. AI has incorporated probability theory to handle uncertainty, utility theory to define objectives, and statistical learning to allow machines to adapt to new circumstances. These developments have created strong connections to other disciplines that build on similar concepts, including control theory, economics, operations research, and statistics.

[2]Gates wrote, "I am in the camp that is concerned about superintelligence. . . . I agree with Elon Musk and some others on this and don't understand why some people are not concerned."

In both the logical-planning and rational-agent views of AI, the machine's objective——whether in the form of a goal, a utility function, or a reward function——is specified exogenously. In Wiener's words, this is "the purpose put into the machine." Indeed, it has been one of the tenets of the field that AI systems should be *general-purpose*——i.e., capable of accepting a purpose as input and then achieving it——rather than *special-purpose*, with their goal implicit in their design. For example, a self-driving car should accept a destination as input instead of having one fixed destination. However, some aspects of the car's "driving purpose" are fixed, such as that it shouldn't hit pedestrians. This is built directly into the car's steering algorithms rather than being explicit: no self-driving car in existence today "knows" that pedestrians prefer not to be run over.

Putting a purpose into a machine that optimizes its behavior according to clearly defined algorithms seems an admirable approach to ensuring that the machine's behavior furthers our own objectives. But, as Wiener warns, we need to put in the *right* purpose. We might call this the King Midas problem: Midas got exactly what he asked for——namely, that everything he touched would turn to gold——but, too late, he discovered the drawbacks of drinking liquid gold and eating solid gold. The technical term for putting in the right purpose is *value alignment*. When it fails, we may inadvertently imbue machines with objectives counter to our own. Tasked with finding a cure for cancer as fast as possible, an AI system might elect to use the entire human population as guinea pigs for its experiments. Asked to de-acidify the oceans, it might use up all the oxygen in the atmosphere as a side effect. This is a common characteristic of systems that optimize: variables not included in the objective may be set to extreme values to help optimize that objective.

Unfortunately, neither AI nor other disciplines built around the optimization of objectives have much to say about how to identify the purposes "we really desire." Instead, they assume that objectives are simply implanted into the machine. AI research, in its present form, studies the ability to achieve objectives, not the design of those objectives. 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.

Steve Omohundro (2008) has pointed to a further difficulty, observing that any sufficiently intelligent entity pursuing a fixed, known objective will act to preserve its own existence (or that of an equivalent successor entity with an identical objective). This tendency has nothing to do with a self-preservation instinct or any other biological notion; it's just that an entity usually cannot achieve its objectives if it's dead. According to Omohundro's argument, a superintelligent machine that has an off-switch——which some, including Alan Turing (1951) himself, have seen as our potential salvation——will take steps to disable the switch in some way. Thus we may face the

prospect of superintelligent machines——their actions by definition unpredictable and their imperfectly specified objectives conflicting with our own——whose motivation to preserve their existence in order to achieve those objectives may be insuperable.

1.3 1001 Reasons to Pay No Attention

Objections have been raised to these arguments, primarily by researchers within the AI community. The objections reflect a natural defensive reaction, coupled perhaps with a lack of imagination about what a superintelligent machine could do. None hold water on closer examination. Here are some of the more common ones:

- *Don't worry, we can just switch it off*:[3] This is often the first thing that pops into a layperson's head when considering risks from superintelligent AI——as if a superintelligent entity would never think of that. It's rather like saying that the risk of losing to Deep Blue or AlphaGo is negligible——all one has to do is make the right moves.
- *Human-level or superhuman AI is impossible*:[4] This is an unusual claim for AI researchers to make, given that, from Turing onward, they have been fending off such claims from philosophers and mathematicians. The claim, which is backed by no evidence, appears to concede that if superintelligent AI were possible, it would be a significant risk. It's as if a bus driver, with all of humanity as his passengers, said, "Yes, I'm driving toward a cliff——in fact, I'm pressing the pedal to the metal. But trust me, we'll run out of gas before we get there." The claim also represents a foolhardy bet against human ingenuity. We've made such bets before and lost. On September 11, 1933, renowned physicist Ernest Rutherford stated, with utter confidence, "Anyone who expects a source of power from the transformation of these atoms is talking moonshine." On September 12, 1933, Leo Szilard invented the neutron-induced nuclear chain reaction. A few years later, he demonstrated such a reaction in his laboratory at Columbia University. As he recalled in a memoir: "We switched everything off and went home. That night, there was very little doubt in my mind that the world was headed for grief."
- *It's too soon to worry about it*: The right time to worry about a potentially serious problem for humanity depends not just on when the problem will occur but also on how much time is needed to devise and implement a solution that avoids the risk. For example, if we were to detect a large asteroid predicted to collide with the Earth in 2070, would we say, "It's too soon to worry!"? And if we consider the global catastrophic risks from climate change predicted to occur later in this century, is it too soon to take action to prevent them? On the contrary, it may be too late. The relevant timescale for human-level AI is less predictable, but, like nuclear fission, it might arrive considerably sooner than expected. Moreover, the technological path to mitigate the risks is also arguably less clear. These two

[3] AI researcher Jeff Hawkins, for example, writes, "Some intelligent machines will be virtual, meaning they will exist and act solely within computer networks. . . . It is always possible to turn off a computer network, even if painful." https://www.recode.net/2015/3/2/11559576/.

[4] The AI100 report (Stone *et al.*, 2016) includes the following assertion: "Unlike in the movies, there is no race of superhuman robots on the horizon or probably even possible."

aspects in combination do not argue for complacency; instead, they suggest the need for hard thinking to occur soon. Wiener (1960) amplifies this point, writing, "The individual scientist must work as a part of a process whose time scale is so long that he himself can only contemplate a very limited sector of it. . . . Even when the individual believes that science contributes to the human ends which he has at heart, his belief needs a continual scanning and re-evaluation which is only partly possible. For the individual scientist, even the partial appraisal of this liaison between the man and the process requires an imaginative forward glance at history which is difficult, exacting, and only limitedly achievable. And if we adhere simply to the creed of the scientist, that an incomplete knowledge of the world and of ourselves is better than no knowledge, we can still by no means always justify the naive assumption that the faster we rush ahead to employ the new powers for action which are opened up to us, the better it will be. We must always exert the full strength of our imagination to examine where the full use of our new modalities may lead us."

One variation on the "too soon to worry about it" argument is Andrew Ng's statement that it's "like worrying about overpopulation on Mars." This appeals to a convenient analogy: Not only is the risk easily managed and far in the future but also it's extremely unlikely that we'd even try to move billions of humans to Mars in the first place. The analogy is a false one, however. We're already devoting huge scientific and technical resources to creating ever more capable AI systems. A more apt analogy would be a plan to move the human race to Mars with no consideration for what we might breathe, drink, or eat once we arrived.

- *It's a real issue but we cannot solve it until we have superintelligence*: One would not propose developing nuclear reactors and *then* developing methods to contain the reaction safely. Indeed, safety should guide how we think about reactor design. It's worth noting that Szilard almost immediately invented and patented a feedback control system for maintaining a nuclear reaction at the subcritical level for power generation, despite having absolutely no idea of which elements and reactions could sustain the fission chain.

  By the same token, had racial and gender bias been anticipated as an issue with statistical learning systems in the 1950s, when linear regression began to be used for all kinds of applications, the analytical approaches that have been developed in recent years could easily have been developed then, and would apply equally well to today's deep learning systems.

  In other words, we can make progress on the basis of general properties of systems–e.g, systems designed within the standard model–without necessarily knowing the details. Moreover, the problem of objective misspecification applies to all AI systems developed within the standard model, not just superintelligent ones.

- *Human-level AI isn't really imminent, in any case*: The AI100 report, for example, assures us, "Contrary to the more fantastic predictions for AI in the popular press, the Study Panel found no cause for concern that AI is an imminent threat to humankind." This argument simply misstates the reasons for concern, which are not predicated on imminence. In his 2014 book, *Superintelligence: Paths, Dangers, Strategies*, Nick Bostrom, for one, writes, "It is no part of the argument in this

book that we are on the threshold of a big breakthrough in artificial intelligence, or that we can predict with any precision when such a development might occur." Bostrom's estimate that superintelligent AI might arrive within this century is roughly consistent with my own, and both are considerably more conservative than those of the typical AI researcher.

- *Any machine intelligent enough to cause trouble will be intelligent enough to have appropriate and altruistic objectives*:[5] This argument is related to Hume's is-ought problem and G. E. Moore's naturalistic fallacy, suggesting that somehow the machine, as a result of its intelligence, will simply perceive what is right given its experience of the world. This is implausible; for example, one cannot perceive, in the design of a chessboard and chess pieces, the goal of checkmate; the same chessboard and pieces can be used for suicide chess, or indeed many other games still to be invented. Put another way: Where Bostrom imagines humans driven extinct by a putative robot that turns the planet into a sea of paperclips, we humans see this outcome as tragic, whereas the iron-eating bacterium *Thiobacillus ferrooxidans* is thrilled. Who's to say the bacterium is wrong? The fact that a machine has been given a fixed objective by humans doesn't mean that it will automatically take on board as additional objectives other things that are important to humans. Maximizing the objective may well cause problems for humans; the machine may recognize those problems as problematic for humans; but, by definition, they are not problematic within the standard model from the point of view of the given objective.
- *Intelligence is multidimensional, "so 'smarter than humans' is a meaningless concept."*: This argument, due to Kevin Kelly (2017), draws on a staple of modern psychology—the fact that a scalar IQ does not do justice to the full range of cognitive skills that humans possess to varying degrees. IQ is indeed a crude measure of human intelligence, but it is utterly meaningless for current AI systems, because their capabilities across different areas are uncorrelated. How do we compare the IQ of Google's search engine, which cannot play chess, with that of Deep Blue, which cannot answer search queries? None of this supports the argument that because intelligence is multifaceted, we can ignore the risk from superintelligent machines. If "smarter than humans" is a meaningless concept, then "smarter than gorillas" is also meaningless, and gorillas therefore have nothing to fear from humans. Clearly, that argument doesn't hold water. Not only is it logically possible for one entity to be more capable than another across all the relevant dimensions of intelligence, it is also possible for one species to represent an existential threat to another even if the former lacks an appreciation for music and literature.

[5] Rodney Brooks (2017), for example, asserts that it's impossible for a program to be "smart enough that it would be able to invent ways to subvert human society to achieve goals set for it by humans, without understanding the ways in which it was causing problems for those same humans." Often, the argument adds the premise that people of greater intelligence tend to have more altruistic objectives, a view that may be related to the self-conception of those making the argument. Chalmers (2010) points to Kant's view that an entity necessarily becomes more moral as it becomes more rational, while noting that nothing in our current understanding of AI supports this view when applied to machines.

1.4 Solutions

Can we tackle Wiener's warning head-on? Can we design AI systems whose purposes don't conflict with ours, so that we're sure to be happy with how they behave? On the face of it, this seems hopeless, because it will doubtless prove infeasible to write down our purposes correctly or imagine all the counterintuitive ways a superintelligent entity might fulfill them.

If we treat superintelligent AI systems as if they were black boxes from outer space, then indeed there is no hope. Instead, the approach we seem obliged to take, if we are to have any confidence in the outcome, is to define some formal problem F and design AI systems to be F-solvers, such that the closer the AI system comes to solving F perfectly, the greater the benefit to humans. In simple terms, the more intelligent the machine, the better the outcome for humans: we hope the machine's intelligence will be applied both to learning our true objectives and to helping us achieve them. If we can work out an appropriate F that has this property, we will be able to create provably beneficial AI. Moreover,

There is, I believe, an approach that may work. Humans can reasonably be described as having (mostly implicit and partially formed) preferences over their future lives——that is, given enough time and unlimited visual aids, a human could express a preference (or indifference) when offered a choice between two future lives laid out before him or her in all their aspects. (This idealization ignores the possibility that our minds are composed of subsystems with effectively incompatible preferences; if true, that would limit a machine's ability to optimally satisfy our preferences, but it doesn't seem to prevent us from designing machines that avoid catastrophic outcomes.) The formal problem F to be solved by the machine in this case is a game-theoretic one: to maximize human future-life preferences subject to its initial uncertainty as to what they are, in an environment that includes human participants. Furthermore, although the future-life preferences are hidden variables, they're grounded in a voluminous source of evidence, namely, all of the human choices ever made.

This formulation sidesteps Wiener's problem, because we do not put a fixed purpose in the machine according to which it can rank all possible futures. Instead, the machine knows that it doesn't know the true preference ranking, so it naturally acts cautiously to avoid violating potentially important but unknown preferences. (We can certainly include fairly strong priors on the positive value of life, health, etc., to make the machine more useful more quickly.) The machine may learn more about human preferences as it goes along, of course, but it will never achieve complete certainty. Such a machine will be motivated to ask questions, to seek permission or additional feedback before undertaking any potentially risky course of action, to defer to human instruction, and to allow itself to be switched off. These behaviors are not built in via preprogrammed scripts or rules; rather, they fall out as solutions of the formal problem F.

As noted in the introduction, this involves a shift from a unary view of AI to a binary one. The classical view, in which a fixed objective is given to the machine, is illustrated qualitatively in Figure 1.1. Once the machine has a fixed objective, it will act to optimize the achievement of the objective; its behavior is effectively independent

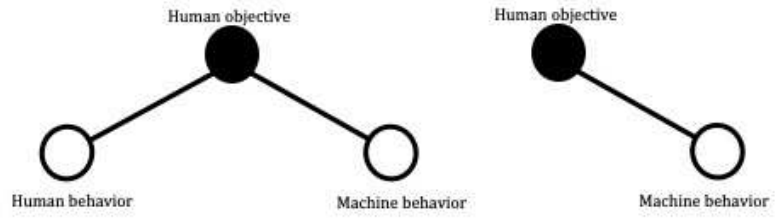

Fig. 1.1 (a) The classical AI situation in which the human objective is considered fixed and known by the machine, depicted as a notional graphical model. Given the objective, the machine's behavior is (roughly speaking) independent of any subsequent human behavior, as depicted in (b). This unary view of AI is tenable only if the human objective can be completely and correctly stated.

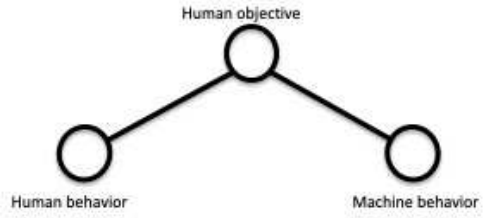

Fig. 1.2 When the human objective is unobserved, machine behavior is no longer independent of human behavior, because the latter provides more information about the human objective.

of the human's behavior.[6]

1.4.1 Assistance games

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. It may involve a form of inverse reinforcement learning (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.

To illustrate assistance games, I'll use the *paperclip game*. It's a very simple game in which Harriet the human has an incentive to "signal" to Robbie the robot some information about her preferences. Robbie is able to interpret that signal because he

[6]The independence is not strict because the human's behavior can provide information about the state of the world. Thus, a passenger in an automated taxi could tell the taxi that snipers have been reported on the road it intends to take, picking off passengers for fun; but this might affect the taxi's behavior only if it already knows that death by gunfire is undesirable for humans.

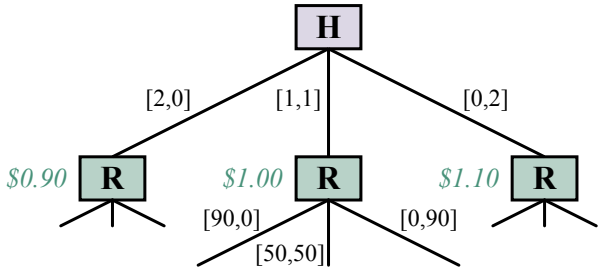

Fig. 1.3 The paperclip game. Each branch is labeled $[p, s]$ denoting the number of paperclips and staples manufactured on that branch. Harriet the human can choose to make two paperclips, two staples, or one of each. (The values in green italics are the values for Harriet if the game ended there, assuming $\theta = 0.45$.) Robbie the robot then has a choice to make 90 paperclips, 90 staples, or 50 of each.

can solve the game and therefore he can understand what would have to be true about Harriet's preferences in order for her to signal in that way.

The steps of the game are depicted in Figure 1.3. It involves making paperclips and staples. Harriet's preferences are expressed by a payoff function that depends on the number of paperclips and the number of staples produced, with a certain "exchange rate" between the two. Harriet's preference parameter θ denotes the relative value (in dollars) of a paperclip; for example, she might value paperclips at $\theta = 0.45$ dollars, which means staples are worth $1 - \theta = 0.55$ dollars. So, if p paperclips and s staples are produced, Harriet's payoff will be $p\theta + s(1 - \theta)$ dollars in all. Robbie's prior is $P(\theta) = Uniform(\theta; 0, 1)$. In the game itself, Harriet goes first, and can choose to make two paperclips, two staples, or one of each. Then Robbie can choose to make 90 paperclips, 90 staples, or 50 of each.

Notice that if she were doing this by herself, Harriet would just make two staples, with a value of $1.10. (See the annotations at the first level of the tree in Figure 1.3.) But Robbie is watching, and he learns from her choice. What exactly does he learn? Well, that depends on how Harriet makes her choice. How does Harriet make her choice? That depends on how Robbie is going to interpret it. One can resolve this circularity by finding a Nash equilibrium. In this case, it is unique and can be found by applying the iterated-best-response algorithm: pick any strategy for Harriet; pick the best strategy for Robbie, given Harriet's strategy; pick the best strategy for Harriet, given Robbie's strategy; and so on. The process unfolds as follows:

1. Start with the greedy strategy for Harriet: make two paperclips if she prefers paperclips; make one of each if she is indifferent; make two staples if she prefers staples.
2. There are three possibilities Robbie has to consider, given this strategy for Harriet:

   (a) If Robbie sees Harriet make two paperclips, he infers that she prefers paperclips, so he now believes the value of a paperclip is uniformly distributed between 0.5 and 1.0, with an average of 0.75. In that case, his best plan is to make 90 paperclips with an expected value of $67.50 for Harriet.

(b) If Robbie sees Harriet make one of each, he infers that she values paperclips and staples at 0.50, so the best choice is to make 50 of each.
(c) If Robbie sees Harriet make two staples, then by the same argument as in (a), he should make 90 staples.

3. Given this strategy for Robbie, Harriet's best strategy is now somewhat different from the greedy strategy in step 1. If Robbie is going to respond to her making one of each by making 50 of each, then she is better off making one of each not just if she is exactly indifferent, but if she is anywhere close to indifferent. In fact, the optimal policy is now to make one of each if she values paperclips anywhere between about 0.446 and 0.554.
4. Given this new strategy for Harriet, Robbie's strategy remains unchanged. For example, if she chooses one of each, he infers that the value of a paperclip is uniformly distributed between 0.446 and 0.554, with an average of 0.50, so the best choice is to make 50 of each. Because Robbie's strategy is the same as in step 2, Harriet's best response will be the same as in step 3, and we have found the equilibrium.

With her strategy, Harriet is, in effect, teaching Robbie about her preferences using a simple code——a language, if you like——that emerges from the equilibrium analysis. Note also that Robbie never learns Harriet's preferences exactly, but he learns enough to act optimally on her behalf——i.e., he acts (given his limited options) just as he would, if he did know her preferences exactly. He is provably beneficial to Harriet under the assumptions stated, and under the assumption that Harriet is playing the game correctly.

It is possible to prove that provided there are no ties that cause coordination problems, finding an optimal strategy for the robot in an assistance game can be done by solving a single-agent partially observable Markov decision process (POMDP) whose state space is the underlying state space of the game plus the human preference parameters θ. POMDPs in general are very hard to solve, but the POMDPs that represent assistance games have additional structure that enables more efficient algorithms (Malik *et al.*, 2018).

1.4.2 The off-switch game

Within the same basic framework, one can also show that a robot solving an assistance game will defer to a human and allow itself to be switched off. This property is illustrated in *off-switch game* shown in Figure 1.4 (Hadfield-Menell *et al.*, 2017*b*). Robbie is now helping Harriet find a hotel room for the Ineternational Paperclip Convention in Geneva. Robbie can act now—let's say he can book Harriet into a very expensive hotel near the meeting venue. He is quite unsure how much Harriet will like the hotel and its price; let's say he has a uniform probability for its net value to Harriet between −40 and +60, with an average of +10. He could also "switch himself off"—less melodramatically, take himself out of the hotel booking process altogether—which is defined (without loss of generality) to have value 0 to Harriet. If those were his two choices, he would go ahead and book the hotel, incurring a significant risk of making Harriet unhappy. (If the range were −60 to +40, with average −10, he would switch himself off instead.) I'll give Robbie a third choice, however: explain his plan, wait,

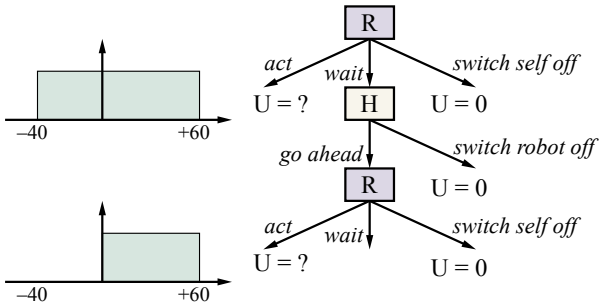

Fig. 1.4 The off-switch game. R, the robot, can choose to act now, with a highly uncertain payoff; to switch itself off; or to defer to H, the human. H can switch R off or let it go ahead. R now has the same choice again. Acting still has an uncertain payoff, but now R knows the payoff is nonnegative.

and let Harriet switch him off. Harriet can either switch him off or let him go ahead and book the hotel. What possible good could this do, one might ask, given that he could make both of those choices himself?

The point is that Harriet's choice—to switch Robbie off or let him go ahead—provides Robbie with information about Harriet's preferences. I'll assume, for now, that Harriet is rational, so if Harriet lets Robbie go ahead, it means the value to Harriet is positive. Now, as shown in Figure 1.4, Robbie's belief changes: it is uniform between 0 and +60, with an average of +30.

So, if one evaluates Robbie's initial choices from his point of view:

1. Acting now and booking the hotel has an expected value of +10.
2. Switching himself off has a value of 0.
3. Waiting and letting Harriet switch him off leads to two possible outcomes:
   (a) There is a 40% chance, based on Robbie's uncertainty about Harriet's preferences, that she will hate the plan and will switch Robbie off, with value 0.
   (b) There is a 60% chance Harriet will like the plan and allow Robbie to go ahead, with expected value +30.

Thus, waiting has expected value $(0.4 \times 0) + (0.6 \times 30) = +18$, which is better than the +10 Robbie epects if he acts now.

The upshot is that Robbie has a positive incentive to defer to Harriet—that is, to allow himself to be switched off. This incentive comes directly from Robbie's uncertainty about Harriet's preferences. Robbie is aware that there's a chance (40% in this example) that he might be about to do something that will make Harriet unhappy, in which case being switched off would be preferable to going ahead. Were Robbie already certain about Harriet's preferences, he would just go ahead and make the decision (or switch himself off); there would be absolutely nothing to be gained from consulting Harriet, because, according to Robbie's definite beliefs, he can already predict exactly what she is going to decide.

In fact, it is possible to prove the same result in the general case: as long as Robbie

is not completely certain that he's about to do what Harriet herself would do, he is better off allowing her to switch him off. Intuitively, her decision provides Robbie with information, and the expected value of information is always nonnegative. Conversely, if Robbie is certain about Harriet's decision, her decision provides no new information, and so Robbie has no incentive to allow her to decide.

Formally, let $P(u)$ be Robbie's prior probability density over Harriet's utility for the proposed action a. Then the value of going ahead with a is

$$EU(a) = \int_{-\infty}^{\infty} P(u) \cdot u \, du = \int_{-\infty}^{0} P(u) \cdot u \, du + \int_{0}^{\infty} P(u) \cdot u \, du \, .$$

On the other hand, the value of action d, deferring to Harriet, is composed of two parts: if $u > 0$ then Harriet lets Robbie go ahead, so the value is u, but if $u < 0$ then Harriet switches Robbie off, so the value is 0:

$$EU(d) = \int_{-\infty}^{0} P(u) \cdot 0 \, du + \int_{0}^{\infty} P(u) \cdot u \, du \, .$$

Comparing the expressions for $EU(a)$ and $EU(d)$, it follows immediately that

$$EU(d) \geq EU(a)$$

because the expression for $EU(d)$ has the negative-utility region zeroed out. The two choices have equal value only when the negative region has zero probability—that is, when Robbie is already certain that Harriet likes the proposed action.

There are some obvious elaborations on the model that are worth exploring immediately. The first elaboration is to impose a cost for Harriet's time. In that case, Robbie is less inclined to bother Harriet if the downside risk is small. This is as it should be. And if Harriet is really grumpy about being interrupted, she shouldn't be too surprised if Robbie occasionally does things she doesn't like.

The second elaboration is to allow for some probability of human error—that is, Harriet might sometimes switch Robbie off even when his proposed action is reasonable, and she might sometimes let Robbie go ahead even when his proposed action is undesirable. It is straightforward to fold this error probability into the model (Hadfield-Menell *et al.*, 2017*b*). As one might expect, the solution shows that Robbie is less inclined to defer to an irrational Harriet who sometimes acts against her own best interests. The more randomly she behaves, the more uncertain Robbie has to be about her preferences before deferring to her. Again, this is as it should be: for example, if Robbie is a self-driving car and Harriet is his naughty two-year-old passenger, Robbie should not allow Harriet to switch him off in the middle of the highway.

The off-switch example suggests some templates for controllable-agent designs and provides a simple example of a provably beneficial system in the sense introduced above. The overall approach 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 we are building one of the agents in order to benefit the other. Unlike a human employee, the robot should have no interests of its own whatsoever.

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, multiple robots, and so on. Scaling up to complex environments and high-dimensional perceptual inputs may be possible using methods related to deep inverse RL. By providing a factored or structured action space, as opposed to the simple atomic actions in the paperclip game, the opportunities for communication can be greatly enhanced. Few of these variations have been explored so far, but I expect the key property of assistance games to remain true: robots that solve such games will be beneficial (in expectation) to humans (Hadfield-Menell *et al.*, 2017*a*).

While the basic theory of assistance games assumes perfectly rational robots that can solve the assistance game exactly, this is unlikely to be possible in practical situations. Indeed, one expects to find qualitatively different phenomena occurring when the robot is much less capable than, roughly as capable as, or much more capable than the human. There is good reason to hope that in all cases improving the robot's capability will be beneficial to the human, because it will do a better job of learning human preferences and a better job of satsifying them.

1.4.3 Acting with unknown preferences

Multiattribute utility theory (Keeney and Raiffa, 1976) views the world as composed of a set of attributes $\{X_1, \ldots, X_n\}$, with preferences defined on lotteries over complete assignments to the attributes. This is clearly an oversimplification, but it suffices for our purpose in exploring some basic phenomena.

In some cases, a machine's scope of action is strictly limited. For example, a (non-Internet-connected) thermostat can only turn the heating on and off, and, to a first approximation, affects the temperature in the house and the owner's bank balance.[7] It is plausible in this case to imagine that the thermostat might develop a decent model of the user's preferences over temperature and cost attributes.

In the great majority of circumstances, however, the AI system's knowledge of human preferences will be extremely incomplete compared to its scope of action. How can it be useful if this is the case? Can it even fetch the coffee?

It turns out that the answer is yes, if we understand "fetch the coffee" the right way. "Fetch the coffee" does not divide the world into goal states (where the human has coffee) and non-goal states. Instead, it says that the human's current preferences rank coffee states above non-coffee states *all other things being equal*. This idea of goals as *ceteris paribus* comparatives is well-established (von Wright, 1972; Wellman and Doyle, 1991). In this context, it suggests that the machine should act in a *minimally invasive* fashion—that is, satisfy the preferences it knows about (coffee) without disturbing any other attributes of the world.

[7] In reality, it is very difficult to limit the effects of the agent's actions to a small set of attributes. Turning the heat off may make the occupants more susceptible to viral infection, and profligate heating may tip the occupants into bankuptcy, and so on. A device connected to the Internet, with the ability to send character streams, can affect the entire planet through propaganda, online trading, etc.

There remains the question of why the machine should assume that leaving other attributes unaffected is better than disturbing them in some random way. One possible answer is some form of risk aversion, but I suspect this is not enough. One has a sense that a machine that does nothing is better than one that acts randomly, but this is certainly not implicit in the standard formulation of MDPs. I think one has to add the assumption that the world is not in an *arbitrary* state; rather, it resembles a state sampled from the stationary distribution that results from the actions of human agents operating according to their preferences (Shah *et al.*, 2019). In that case, one expects a random action to make things worse for the humans.

There is another kind of action that is beneficial to humans even when the machine knows nothing at all about human preferences: an action that simply expands the set of actions available to the human. For example, if Harriet has forgotten her password, Robbie can give her the password, enabling a wider range of actions than Harriet could otherwise execute.

1.5 Reasons for Optimism

There are some reasons to think this approach may work in practice. First, there is abundant written and filmed information about humans doing things (and other humans reacting). More or less every book ever written contains evidence on this topic. Even the oldest clay tablets, tediously recording the exchange of N sheep for M oxen, give information about human preferences between sheep and oxen. Technology to build models of human preferences from this storehouse will presumably be available long before superintelligent AI systems are created.

Second, there are strong near-term economic incentives for robots to understand human preferences, which also come into play well before the arrival of superintelligence. Already, computer systems record one's preferences for an aisle seat or a vegetarian meal. More sophisticated personal assistants will need to understand their user's preferences for cost, luxury, and convenient location when booking hotels, and how these preferences depend on the nature and schedule of the user's planned activities. Managing a busy person's calendar and screening calls and emails requires an even more sophisticated understanding of the user's life, as does the management of an entire household when entrusted to a domestic robot. For all such roles, trust is essential but easily lost if the machine reveals itself to lack a basic understanding of human preferences. If one poorly designed domestic robot cooks the cat for dinner, not realizing that its sentimental value outweighs its nutritional value, the domestic-robot industry will be out of business.

For companies and governments to adopt the new model of AI, a great deal of research must be done to replace the entire toolbox of AI methods, all of which have been developed on the assumption that the objective is known exactly. There are two primary issues for each class of task environments: how to relax the assumption of a known objective and what form of interaction to assume between the machine and the human. For example, problem-solving task environments have an objective defined by a goal test $G(s)$ and a stepwise cost function $c(s, a, s')$. Perhaps the machine knows a relaxed predicate $G' \supset G$ and upper and lower bounds c^+ and c^- on the cost function, and can ask the human (1) whether any given state s satisfies G and (2) whether one

trajectory to s is preferred to another. Design considerations include formal precision, algorithmic complexity, feasibility of the interaction protocol from the human point of view, and applicability in real-world circumstances.

The standard model of AI as maximizing objectives does not imply that all AI systems have to solve some particular problem formulation such as an influence diagram or a factored MDP. For example, it is entirely consistent with the standard model to build an AI system directly as a policy, expressed as a set of condition–action rules specifying the optimal action in each category of states.[8] By the same token, I am not proposing that all AI systems under the new model have to solve some explicitly formulated representation of the assistance game. It is important to maintain a broad conception of the approach and how it applies to the design of AI systems for any particular task environment. The crucial elements are (1) acknowledgement that there is partial and uncertain information about the true human preferences that are relevant in the task environment; (2) a means for information to flow at run-time from humans to machines concerning those preferences; and (3) allowance for the human to be a joint participant in the run-time process.

1.6 Obstacles

There are obvious difficulties with an approach that expects machines to learn underlying preferences from observing human behavior. The first 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 decision problems we face. For example, if two humans are playing chess and one of them loses, it's 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.

As yet, we do not know enough about human cognitive mechanisms to invert real human behavior to get at the underlying preferences. One thing that seems intuitively clear, however, is that one of our principal methods for coping with the complexity of the world is 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 would do, we choose among subtasks such as "write the introduction" 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

[8] Many applications of control theory work exactly this way: the control theorist works offline with a mathematical model of the system and the objective to derive a *control law* that is then implemented in the controller.

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.

Machines might try to discover more about human cognitive mechanisms by an inductive learning approach. Suppose that in some given state s Harriet's action a depends on her preferences θ according to mechanism h, i.e., $a = h(\theta, s)$. (Here, θ represents not a single parameter such as the exchange rate between staples and paperclips, but the Harriet's preferences over future lives, which could be a structure of arbitrary complexity.) By observing many examples of s and a, is it possible eventually to recover h and θ? At first glance, the answer seems to be no (Armstrong and Mindermann, 2019). For example, one cannot distinguish between the following hypotheses about how Harriet plays chess:

1. h maximizes the satisfaction of preferences, and θ is the desire to win games.
2. h minimizes the satisfaction of preferences, and θ is the desire to lose games.

From the outside, Harriet plays perfect chess under either hypothesis.[9] If one is merely concerned with predicting her next move, it doesn't matter which formulation one chooses. On the other hand, for a machine whose goal is to help Harriet realize her preferences, it really does matter! The machine needs to know which explanation holds. From this viewpoint, something is seriously wrong with the second explanation of behavior. If Harriet's cognitive mechanism h were really trying to minimize the satisfaction of preferences θ, it wouldn't make sense to call θ her preferences. It is, then, simply a mistake to suppose that h and θ are separately and independently defined. I have already argued that the assumption of perfect rationality——i.e., h is maximization——is too strong; yet, for it to make sense to say that Harriet has preferences, h will have to satisfy (or nearly satisfy) some basic properties associated with rationality. These might include choosing correctly according to preferences in situations that are computationally trivial——for example, choosing between vanilla and bubble-gum ice cream at the beach. Cherniak (1986) presents an in-depth analysis of these minimal conditions on rationality.

Further difficulties arise if the machine succeeds in identifying Harriet's preferences, but finds them to be inconsistent. For example, suppose she prefers vanilla to bubble gum and bubble gum to pistachio, but prefers pistachio to vanilla. In that case her preferences violate the axiom of transitivity and there is no way to maximally satisfy her preferences. (That is, whatever ice cream the machine gives her, there is always another that she would prefer.) In such cases, the machine could attempt to satisfy Harriet's preferences *up to inconsistency*; for example, if Harriet strictly prefers all three of the aforementioned flavors to licorice, then it should avoid giving her licorice ice cream.

Of course, the inconsistency in Harriet's preferences could be of a far more radical nature. Many theories of cognition, such as Minsky's *Society of Mind* (1986), posit multiple cognitive subsystems that, in essence, have their own preference structures

[9] Of course, the Harriet who prefers to lose might grumble when she keeps winning, thereby giving a clue as to which Harriet she is. One response to this is that grumbling is just more behavior, and equally subject to multiple interpretations. Another response is to say that Harriet might feel grumbly but, in keeping with her minimizing h, would instead jump for joy. This is not to say that there is no fact of the matter as to whether Harriet is pleased or displeased with the outcome.

and compete for control——and these seem to be manifested in addictive and self-destructive behaviors, among others. Such inconsistencies place limits on the extent to which the idea of machines helping humans even makes sense.

Also difficult, from a philosophical viewpoint, is the apparent *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, because 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 Harriet the machine should try to help: Harriet_{2020}, Harriet_{2035}, or some time-averaged Harriet? (See Pettigrew (2020) for a full treatment of this approach, wherein decisions for individuals who change over time are made as if they were decisions made on behalf of multiple distinct individuals.) Plasticity is also problematic because of the possibility that the machine may, by subtly influencing Harriet's environment, gradually mould her preferences in directions that make them easier to satisfy, much as certain political forces have been said to do with voters in recent decades.

I am often asked, "Whose values should we align AI with?" (The question is usually posed in more accusatory language, as if my secret, Silicon-Valley-hatched plan is to align all the world's AI systems with my own white, male, Western, cisgender, Episcopalian values.) Of course, this is simply a misunderstanding. The kind of AI system proposed here is not "aligned" with any values, unless you count the basic principle of helping humans realize their preferences. For each of the billions of humans on Earth, the machine should be able to predict, to the extent that its information allows, which life that person would prefer.

Now, practical and social constraints will prevent all preferences from being maximally satisfied simultaneously. We cannot all be Ruler of the Universe. This means that machines must mediate among conflicting preferences——something that philosophers and social scientists have struggled with for millennia. At one extreme, each machine could pay attention only to the preferences of its owner, subject to legal constraints on its actions. This seems undesirable, as it would have a machine belonging to a misanthrope refuse to aid a severely injured pedestrian so that it can bring the newspaper home more quickly. Moreover, we might find ourselves needing many more laws as machines satisfy their owners' preferences in ways that are very annoying to others even if not strictly illegal. At the other extreme, if machines consider equally the preferences of all humans, they might focus a larger fraction of their energies on the least fortunate than their their owners might prefer—a state of affairs not conducive to investment in AI. Presumably, some middle ground can be found, perhaps combining a degree of obligation to the machine's owner with public subsidies that support contributions to the greater good. Determining the ideal solution for this issue

is an open problem.

Another common question is, "What if machines learn from evil people?" Here, there is a real issue. It is *not* that machines will learn to copy evil actions. The machine's actions need not resemble in any way the actions of those it observes, any more than a criminologist's actions resemble those of the criminals she observes. The machine is learning about human preferences; it is not adopting those preferences as its own and acting to satisfy them. For example, suppose that a corrupt passport official in a developing country insists on a bribe for every transaction, so that he can afford to pay for his children to go to school. A machine observing this will not learn to take bribes itself: it has no need of money and understands (and wishes to avoid) the toll imposed on others by the taking of bribes. The machine will instead find other, socially beneficial ways to help send the children to school. Similarly, a machine observing humans killing each other in war will not learn that killing is good: obviously, those on the receiving end very much prefer not to be dead.

The difficult issue that remains is this: what should machines learn from humans who enjoy the suffering of others? In such cases, any simple aggregation scheme for preferences (such as adding utilities) would lead to some reduction in the utilities of others in order to satisfy, at least partially, these perverse preferences. It seems reasonable to require that machines simply ignore positive weights in the preferences of some for the suffering of others (Harsanyi, 1977).

1.7 Looking Further Ahead

If we assume, for the sake of argument, that all of these obstacles can be overcome, as well as all of the obstacles to the development of truly capable AI systems, are we 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 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 noted earlier are of the kind that 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. We already see this in miniature with the imposition of machine-checkable software standards for cell-phone applications.

On the question of enforcement of policies for AI software design, I am less sanguine. If Dr. Evil wants to take over the world, he or she might remove the safety catch, so to speak, and deploy an AI system that ends up destroying the world instead. This problem 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 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 just a *little* more helpful than they would otherwise wish to be. Unfortunately, this process continues in a direction that is hard to resist.

1.8 Conclusion

Finding a solution to the AI control problem is an important task; it may be, in Bostrom's words, "the essential task of our age." It involves building systems that are far more powerful than ourselves while still guaranteeing that those systems will remain powerless, forever.

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. No matter how excellently an algorithm maximizes, and no matter how accurate its model of the world, a machine's decisions may be ineffably stupid, in the eyes of an ordinary human, if it fails to understand human preferences.

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 to it.

References

Armstrong, Stuart and Mindermann, Sören (2019). Occam's razor is insufficient to infer the preferences of irrational agents. In *Advances in Neural Information Processing Systems 31*.

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

Brooks, Rodney (2017). The seven deadly sins of AI predictions. MIT Technology Review, October 6.

Chalmers, David John (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies*, **17**, 7–65.

Chan, Lawrence, Hadfield-Menell, Dylan, Srinivasa, Siddhartha, and Dragan, Anca (2019). The assistive multi-armed bandit. In *Proceedings of the Fourteenth ACM/IEEE International Conference on Human–Robot Interaction*.

Cherniak, C. (1986). *Minimal Rationality*. MIT Press.

Gates, W. (2015). Ask me anything. Reddit, January 28.

Hadfield-Menell, Dylan, Dragan, Anca D., Abbeel, Pieter, and Russell, Stuart J. (2017*a*). Cooperative inverse reinforcement learning. In *Advances in Neural Information Processing Systems 29*.

Hadfield-Menell, Dylan, Dragan, Anca D., Abbeel, Pieter, and Russell, Stuart J. (2017*b*). The off-switch game. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*.

Harsanyi, John (1977). Morality and the theory of rational behavior. *Social Research*, **44**, 623–656.

Keeney, Ralph L. and Raiffa, Howard (1976). *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley.

Kelly, Kevin (2017). The myth of a superhuman AI. *Wired*, April 25.

Kumparak, G. (2014). Elon Musk compares building artificial intelligence to 'summoning the demon'. TechCrunch, October 26.

Malik, Dhruv, Palaniappan, Malayandi, Fisac, Jaime F., Hadfield-Menell, Dylan, Russell, Stuart J., and Dragan, Anca D. (2018). An efficient, generalized Bellman update for cooperative inverse reinforcement learning. In *Proceedings of the Thirty-Fifth International Conference on Machine Learning*.

Minsky, Marvin L. (1986). *The Society of Mind*. Simon and Schuster.

Ng, Andrew Y. and Russell, Stuart J. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*.

Omohundro, S. (2008). The basic AI drives. In *AGI-08 Workshop on the Sociocultural, Ethical and Futurological Implications of Artificial Intelligence*.

Osborne, H. (2017). Stephen Hawking AI warning: Artificial intelligence could destroy

civilization. Newsweek, November 7.

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

Russell, Stuart J. (1998). Learning agents for uncertain environments. In *Proceedings of the Eleventh ACM Conference on Computational Learning Theory*.

Shah, Rohin, Krasheninnikov, Dmitrii, Alexander, Jordan, Abbeel, Pieter, and Dragan, Anca (2019). The implicit preference information in an initial state. In *Proceedings of the Seventh International Conference on Learning Representations*.

Stone, Peter, Brooks, Rodney A., Brynjolfsson, Erik, Calo, Ryan, Etzioni, Oren, Hager, Greg, Hirschberg, Julia, Kalyanakrishnan, Shivaram, Kamar, Ece, Kraus, Sarit et al. (2016). Artificial intelligence and life in 2030. Technical report, Stanford University One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel.

Turing, A. (1951). Can digital machines think? Lecture broadcast on BBC Third Programme. Typescript available at www.turingarchive.org.

von Wright, Georg (1972). The logic of preference reconsidered. *Theory and Decision*, **3**, 140–167.

Wellman, M. P. and Doyle, Jon (1991). Preferential semantics for goals. In *Proceedings of the Ninth National Conference on Artificial Intelligence*.

Wiener, Norbert (1960). Some moral and technical consequences of automation. *Science*, **131**(3410), 1355–1358.