Humanity Compatible: Aligning Autonomous AI with Kantian Respect for Humanity

Ava Thomas Wright, California Polytechnic State University, San Luis Obispo
International Conference on Computer Ethics: Philosophical Enquiry (CEPE) 2023, Chicago, IL

Keywords: Kant, Kantian autonomy, AI autonomy, machine ethics, value alignment, Stuart Russell, rational agency

Extended Abstract

Artificially intelligent autonomous agents are "autonomous" in the sense that they are programmed to learn for themselves how to act across a wide range of situations. Autonomous agents are programmed with a set of objectives, rewards and punishments, constraints, or other performance measures, and then progressively learn behaviors to optimize or satisfy those measures. What they will do in any given situation, therefore, cannot be completely foreseen or predicted in advance.

Computer scientist Stuart Russell criticizes the "standard model" in autonomous AI agent design in which agents are programmed to optimize their behavior to achieve fixed objectives that we specify for them.[1] The problem with this model is that a highly capable autonomous AI agent might pursue those objectives in unforeseen ways that violate our values and outstrip our control. Russell argues that AI agents should be designed, instead, to *learn what our objectives are*, while necessarily remaining uncertain about them.

> We don't want machines that are intelligent in the sense of pursuing their objectives; we want them to pursue our objectives. If we cannot transfer those objectives perfectly to the machine, then we need a new formulation – one in which the machine is pursuing our objectives, but is necessarily uncertain as to what they are. (Russell and Norvig, 2020)

An advanced autonomous AI designed to pursue our objectives could never be certain what they are; therefore, Russell argues, the agent would always have an incentive to seek guidance from us before acting. The more uncertain the agent is, the more such incentive it would have. On Russell's model, autonomous AI thus would be designed to align its behavior with our values from the start. Russell refers to autonomous AI agents designed on this new model as "human compatible" AI.

---

[1] See Russell, S. *Human Compatible* (2018); see also Russell, S. and P. Norvig. *Artificial Intelligence: A Modern Approach* (AIMA), 4th ed. (2021), which is the standard textbook used in most introductory AI courses in computer science. Russell and Norvig revised the 2010 third edition to reflect the views on human rational agency and value alignment that Russell sets out in *Human Compatible*.

Russell's new model for the design of autonomous AI is an important innovation. But the model raises a new question, How do we program an autonomous AI to learn what our objectives are from its observations of what we say or do? What assumptions should it bring to its task of discovering our objectives from observations of what we say and do?

According to the view Russell favors, human objectives are best understood as the satisfaction of preferences over complete future lives. This approach is inspired by a model of rational human agency advanced by economists, who have shown that so long as our preference rankings obey various constraints – completeness, transitivity, etc. – our behavior can be understood as the attempt to maximize a utility function. Russell's autonomous AI would thus attempt to reconstruct this utility function from its observations of our behavior, so that it could then help us to maximize it.

But human rational agency is not limited to finding instrumentally efficient means to maximize the satisfaction of preferences presumably given to us by our natural inclinations. Humans also may have *substantive* reasons to set certain ends such as helping others, or to do or avoid certain actions such as committing murder or breaking an important promise. Unlike other animals (or, indeed, machines), we are morally responsible for our choices, which implies that we can act for moral as well as strictly instrumental reasons. Immanuel Kant refers to our rational capacity to act out of respect for moral law as our "humanity," and it is the foundation of our moral and legal rights.

In this paper, I will argue that autonomous AI agents designed along Russellian lines should be programmed to determine our objectives by modeling our agency substantively as Kantian autonomy, rather than as the satisfaction of preferences. The objectives that the AI infers from our behavior should not be understood in terms of efforts to satisfy preferences but instead in terms of efforts to act autonomously in the Kantian sense of that term.[2] Only by modeling us as autonomous agents will the AI be able to learn and help us to achieve our objectives. Any other model of our agency would fail to treat us as ends and so fail to respect our humanity. I thus argue for *humanity* compatible AI.

Consider a case in which I decline to undergo some medical treatment I need because I am afraid of the pain, and let's stipulate that undergoing the treatment serves my overall good.[3] It seems like the only way the economic account can make sense of my decision is to say that I am making a mistake, and that I would automatically change my mind if I had better information. Kantians think it is more plausible to say that I am acting irrationally: I know

---

[2] AI "autonomy" is quite different from Kantian autonomy or freedom. While autonomous machine agents might be understood to have various "incentives" for action oriented toward achieving competing performance measures, they are not capable of *freely choosing for themselves* which such incentives to take as their motivating reason for action. They will always act in accordance with whatever incentives best optimize or satisfy their performance measures in the ways that they have been programmed. Thus while autonomous machine agents can be programmed to do what is right, they cannot be programmed to freely choose to do what is right for the reason that it is right, which is what Kantian autonomy requires (see G: 4:397)

[3] See Korsgaard (2008), The Normativity of Practical Reason. In: The Constitution of Agency, p. 10.

that rationally I ought to undergo the treatment—that doing so would promote my utility in terms of preference rankings over future lives--but I am giving in to my fear. People fail to do what they know they rationally should all the time, perhaps because they are too lazy or depressed. (The economic theory cannot say that I am acting irrationally without abandoning the instrumental principle.)

It seems like there is some danger that the preference-satisfying AI agent would assume that what I "really want" is the treatment (given that we have stipulated that I know it to be my overall good) and then force or manipulate me undergo it. Or, if I had strong preferences against such manipulation, It would at best treat me like a child who refuses to take her medicine. A Kantian AI agent, by contrast, will assume that I have willed the end of my own happiness and that I need the treatment in order to promote it, but the agent would not assume that I have "really" willed the treatment (even setting aside that manipulating me would undermine my autonomy). It would first of all respect my will, while secondarily helping me to overcome my fears. The preference-satisfying agent seems to have this backwards. It would help me to overcome my fears so that I would take the treatment without protest, but it would regard my choice to undergo it as valuable only as a means to promoting my overall good. It would, in fact, treat my rational agency merely as a means toward its ends.

One of my deep worries about Russell's preference-satisfaction utilitarian approach to building human-compatible AI is that such agents would conceive human rational agency as no different from the agency of other animals. Over the long term, the influence of powerful AI agents treating us this way would undermine human dignity. They may help us to become happy animals, but we should strive to be better.