Toward Substantive Models of Rational Agency in the Design of Autonomous AI

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

Keywords: autonomous machine agent, AI, value alignment, control, rational agency, machine ethics, substantive reason, instrumental reason

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. AI autonomy gives rise to the problem of *value alignment*, How can we make sure that AI agents, acting autonomously, will behave in ways that align with moral values? The related *control problem* is, How can we make sure that autonomous AI agents that may be more intelligent and capable than us will remain under our control? The control problem is related to value alignment because the worry is that we may not be able to control whether powerful AI agents of the future act in ways that align with moral values.

Computer scientist Stuart Russell argues that these problems arise because of a fundamental flaw in what he calls the "standard model" in AI agent design. On the standard model, AI agents are programmed to optimize their behavior to achieve objectives that we specify for them. Russell argues that AI agents should be designed, instead, to try to determine what our objectives are. Autonomous AI agents will then always have an incentive to seek guidance from us about our objectives before acting upon them, which will ensure that their behavior remains aligned with our values, Russell argues.

In this paper, we will argue that these problems cannot be solved so long as AI rational agency is conceived strictly instrumentally. A more substantive conception of rational agency is needed, one in which autonomous machine agents reason not only about how to efficiently achieve their ends, but also about what those ends should be. In arguing for this position, we will examine recent work by computer scientist Stuart Russell.[1] Russell's view represents an important retreat from purely instrumental notions of rational agency. But we think Russell should have gone

further.

Russell takes issue with what he calls the "standard model" in AI—the idea that autonomous machines should be designed to optimize on fixed objectives. The standard model faces what Russell calls the King Midas Problem: it is difficult to specify our objectives for an autonomous machine in a way that is sufficiently precise to avoid unintended and potentially harmful consequences. Russell says that instead of optimizing on fixed objectives that we specify, autonomous machine agents should attempt to discover what our objectives are. He writes,

> 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.[1]

The standard model represents an extreme form of instrumentalism, where intelligent machines find efficient means to whatever ends we specify. Russell's idea is to reconceive the goal of AI research as creating machines that attempt to determine what our ends are, so they can assist us in achieving them. We think that this is a helpful development in the direction of a more substantive notion of intelligence for artificial autonomous agents. For Russell, intelligent machines reason about what ends they should pursue by trying to figure out what human beings want. But, as Russell himself recognizes, it is easy to imagine cases where doing what human beings want is not an especially good thing to do. We think it important that autonomous machine agents can make that assessment—about what ends are good—for themselves. Another way to put this idea is to say that machine agents should be responsive to *normative reasons* to adopt ends.

A machine that took sophisticated and efficient means to pointless ends would be called intelligent by proponents of the standard model. If humans wanted those pointless ends to be realized, then Russell would call that machine "intelligent" as well. But we think no matter how sophisticated it is, no matter how good at satisfying human preferences, if a machine doesn't do what there is good normative reason to do, then it isn't really an intelligent machine. Building such machines does not necessarily constitute progress in AI research.

A core part of Russell's approach to value alignment and control is the idea that autonomous machine agents need to remain *uncertain* as to what our ends really are. Agents that remain

[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. In this major new edition of AIMA, Russell and Norvig revise the 2010 third edition to reflect the views on rational agency, value alignment, and control that Russell set out in *Human*

Compatible.
uncertain about our ends will always have some incentive to seek guidance from us—including permitting themselves to be switched off—before taking action to try to achieve those ends. Russell's argument for the importance of uncertainty proceeds via an analysis of a coordination game called The Off Switch, the details of which we will examine in the full paper. We agree [2] with Russell that uncertainty about our ends is essential, but it plays an ad hoc role in Russell's account, and it isn't clear exactly what kind of uncertainty is required. If, instead, we conceive the goal of intelligent machines as reflecting on and being responsive to normative reasons, then we'll be able to explain both why uncertainty about our ends is an important part of intelligence and exactly what kind of uncertainty is needed.

We do not go so far as to argue that AI agents should be autonomous in the Kantian sense. Machine agents are not now and may never be capable of the freedom required for Kantian autonomy and virtue. Our aims are more modest. We believe that like many economists, psychologists and legal theorists in the 20th century, AI researchers have proposed objectionably reductive analyses of normative concepts and problematically instrumental conceptions of rationality. We aim to promote a shift toward the design of machines that are rational in the sense of reflecting on and responding appropriately to normative reasons. We suggest some new research directions by making analogies with how other fields have made a similar shift.

[2][3] See Russell (2018), chap. 8.