HUMAN-CENTERED AI: THE ARISTOTELIAN APPROACH

JACOB SPARKS[*] – AVA THOMAS WRIGHT[**]

THE HUMAN IN HUMAN-CENTERED AI

Suppose humanity succeeds in building an artificial general intelligence (AGI), an AI capable of finding adequate means to achieve almost any given objective. We want the AI to help us, but to do so, we must tell it how. What objective should we give it?

If you spend any time thinking about this question, you're liable to remember the many myths and stories that caution against careless wishing. King Midas got what he wished for, but it didn't turn out well for him. In a similar vein, we should be careful about selecting the aim for our AI. We might not like to get exactly what we ask for. As Norbert Wiener warns:

> 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 the purpose which we really desire[1].

Call this the *specification problem*: it's difficult to specify the objectives of an AI in a way that anticipates everything that could go

[*] *CalPoly*, California, USA.

[**] *CalPoly*, California, USA.

[1] N. WIENER, *Some Moral and Technical Consequences of Automation*, in «Science» 131, 3410 (1960), pp. 1355-1358.

wrong[2]. The specification problem is a problem for any system that operates autonomously, but it becomes increasingly challenging as our machines operate autonomously across broader contexts. It reaches its most difficult form as we imagine a general intelligence that is meant to assist us with a wide range of our activities.

One increasingly popular solution to the specification problem, which we'll call *human-centered*, involves tying the machine's objectives to our objectives, i.e. giving the machine the task of learning our objectives and helping us to satisfy them[3]. We are spared the trouble of specifying our objectives, since part of the machine's objective is to learn them. In their widely used textbook, Stuart Russell and Peter Norvig write:

> 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[4].

Though human-centered approaches obviate the need to specify our objectives, they require us to structure the learning problem for the machine. How should it go about learning our objectives? What model of human agency should it bring to its observations of our behavior as it seeks to discover our true objectives?

[2] This problem is closely related to the "value alignment problem", i.e. the problem of creating advanced AI that behaves in ways consistent with our values. We are highlighting the fact that one approach to alignment – specification of our values – comes with serious difficulties.

[3] See, e.g., E. YUDKOWSKY, *Coherent Extrapolated Volitions*, The Singularity Institute, San Francisco 2001; N. BOSTROM, *Superintelligence: Paths, dangers, strategies*, Oxford University Press, New York 2014; S. J. RUSSELL, *Human compatible: artificial intelligence and the problem of control*, Viking, New York 2019.

[4] S. RUSSELL, P. NORVIG, *Artificial Intelligence: A Modern Approach*, 4th ed., Pearson, Hoboken, N.J. 2021, p. 5.

Russell and Norvig's own answer draws from theories of human rational agency familiar from economics[5]. We are understood to have preferences for all the different possible future lives we might live, where those preferences obey various axioms to secure their coherence. For example, the transitivity axiom enforces the rational requirement that if we prefer future life A over B, and future B over C, then we should also prefer A over C. Our behavior is then conceived as an attempt to maximize a utility function defined by these coherent preferences, which the economic model also takes to be roughly stable. On this model, an AI assistant would observe our behavior, learn our preference rankings over possible future lives, construct our utility functions, and help us maximize them.

The question of how human-centered AI conceives human agency is crucial. Given the long history of criticism of the economic model, we think it's important to explore alternatives[6]. In this paper, we contrast the economic conception of our rational agency with one derived from Aristotle[7]. These are not the only models of human rationality

[5] See S. RUSSELL, *Human compatible: artificial intelligence and the problem of control*, cit., p. 176, where he clarifies that the human objectives he wants machines to learn about and satisfy should be understood in the way economists understand "preference" and "utility".

[6] From the social sciences, there is the *behaviorist* critique, exemplified by H. SIMON, *Models of Man*, John Willey, New York 1957, and D. KAHNEMAN, A. TVERSKY, *Prospect Theory: An Analysis of Decision under Risk*, in «Econometrica» 47, 2 (1979), pp. 263-291, who claim the traditional economic models are predictively inaccurate. From philosophy there are various *normative* criticisms, e.g. E. ANDERSON, *Value in Ethics and Economics*, Harvard University Press, Cambridge 1993, and M. C. NUSSBAUM, *Flawed Foundations: The Philosophical Critique of (A Particular Type of) Economics*, in «The University of Chicago Law Review» 64, 4 (1997), pp. 1197-1214, that claim the economic model makes false assumptions about our normative reasons for action.

[7] We are not here trying to give a definitive interpretation of Aristotle. Our aim is, instead, to outline a model of rational agency inspired by Aristotle that serves both as a contrast to the economic model and as a guide to researchers building human-centered AI.

available. This paper is part of a larger project exploring the consequences of various ways of conceiving human rationality as we build human-centered AI. What we argue here is that aspects of the Aristotelian model are worth taking seriously if we want to build truly beneficial human-centered AI.

The economic and Aristotelian models, as applied to the problem of learning our objectives from our behavior, are also not mutually exclusive positions. There are many variations on each approach and many hybrid possibilities. So rather than suggest that the Aristotelian model is superior, we take our arguments to show only that researchers, as they build increasingly intelligent machines, should respect aspects of our rational agency that the economic model risks omitting.

Economic and aristotelian models

The economic model of rational agency is *structural*. Rational agents have stable, coherent preferences and they intend the means to their ends. It is thus theoretically possible, on structural views, to rationally pursue immoral or imprudent ends, as long as doing so structurally coheres with all your other ends. Aristotle, by contrast, has a substantive model. Rationality is about selecting appropriate ends as well as appropriate means toward those ends.

Some, including Russell, take Aristotle to support structural views when he writes,

> We deliberate not about ends, but about means. For a doctor does not deliberate whether he shall heal, nor an orator whether he shall persuade […] They assume the end and consider how and by what means it is attained […][8].

If Aristotle thinks we deliberate only about means, then that suggests that rationality doesn't set any requirements on our ends. But Aristotle

[8] Aristotle, *Nicomachean Ethics*, trans. T. Irwin, Hackett Publishing Company, Indianapolis, 1999, 1112b11-12.

would not agree that it is possible to behave rationally in the service of imprudent or immoral ends. He writes,

> The unqualifiedly good deliberator is the one whose aim accords with rational calculation in pursuit of the best good for a human being that is achievable in action[9].

And elsewhere, Aristotle notes,

> It seems proper to a [wise] person to be able to deliberate finely about things that are good and beneficial for himself…about what sorts of things promote living well in general[10].

Hence, according to Aristotle, rational deliberation consists in reasoning well about the means to achieve *good*, not bad ends, which implies some degree of reasoning about what ends are good or bad. What does Aristotle mean, then, when he says that we "deliberate not about ends, but means"?

The answer is that Aristotle sees deliberation as involving insight into what actions *constitute* the end, rather than what actions will cause the end, as on the economic model. Aristotle draws an analogy between practical and geometrical reasoning[11]. When a geometer tries to construct, for example, a square inscribed in a circle, she doesn't look for just any means to create the desired figure, but only those means that construct it while preserving the figure's status *as a geometrical construction*. If the desired figure isn't a geometrical construction to begin with, then there is no possibility of finding those means[12]. Likewise, if your end isn't itself good, there is no possibility of deliberating well in your attempt to attain it.

[9] *Ivi*, 1141b12-14.

[10] *Ivi*, 1140a25-28.

[11] *Ivi*, 1112b20-4.

[12] See A. CALLARD, «Aristotle on deliberation», in R. CHANG, K. SYLVAN, eds., *Routledge Handbook of Practical Reason*, Routledge, New York 2020, pp. 126-140.

The economic and Aristotelian models accordingly differ in the language they use to describe human ends. While the economic model represents our preferences in descriptive language, Aristotle has a "thick" conception of our ends. You might call the doctor's true end, "good health", and the orator's end, "good persuasion", though the "good" in those phrases would be redundant for Aristotle. You can't achieve "good persuasion" if you are trying to get your audience to believe something false, or if you use manipulative means, or if you sacrifice other goods that constitute a good human life. Deliberating well in the service of bad ends is thus impossible for Aristotle.

Another difference between the economic and Aristotelian models is that the latter emphasizes that having certain virtues of character is necessary both to reason well about what one should do in some situation and then to actually do it. First, to determine what one should do in some situation, one must be able to perceive what features of the situation are relevant, which is difficult to do without the right character dispositions. For example, a dull or callous person may not perceive that someone recently bereaved is grieving and so needs comforting, whereas a more sensitive or compassionate person would perceive that grief. Then, to actually do what one rationally should in the situation, one must do it in the right way, with the right affect, which may also be a function of one's character. Even if the dull or callous person correctly perceives that someone needs comforting, if they feel no sympathy for the bereft, then any attempt at comfort likely will be clumsy and ineffective.

Aristotle emphasizes the dispositional and affective aspects of rational action in his description of the virtue of courage,

> Hence whoever stands firm against the right things, and fears the right things, for the right end, in the right way, at the right time, and is correspondingly confident, is the courageous person; for the courageous person's actions and feelings accord with what something is worth, and follow what reason prescribes[13].

[13] *Ivi*, 1115b16-19.

Achilles does what he should on the battlefield because his character exemplifies the right balance between rashness and cowardice, and he feels the right amount of fear and anger at the right things. These traits of character are necessary for him to perceive and then do what is right. Because the economic model more or less ignores the dispositional and affective aspects of practical rationality, it has trouble capturing some of the ways that Aristotle thinks our behavior might not express our true objectives.

FAILURES OF RATIONALITY

Both the economic and Aristotelian views agree that our behavior doesn't always reflect our true objectives, but they disagree about the kinds of irrationality that can explain the divergence. As a result, they have different approaches to the problem of discovering and helping us achieve our objectives. In this section, we look at cases illustrating forms of irrationality and investigate these divergent approaches.

We want to emphasize at the outset that learning our objectives and helping us pursue them are two distinct tasks. The first is an interpretive activity, the second is an assistive one. Both are affected by the model of human rationality you bring to the task of building human-centered AI.

Bad ends

The idea that we can be irrational by failing to have good ends is a significant departure from the economic model. Consider Ebenezer, who has developed an unhealthy obsession with money. He works tirelessly, disregarding his relationships and other interests. Nevertheless, his behavior is perfectly consistent with his preference for money over these other goods.

From the perspective of the economic model, it seems the AI should assist Ebenezer in pursuing money. His preference appears to be for a future life with riches, regardless of the damage this might cause to his other interests. The economic AI would need to carefully check to make sure that Ebenezer really is consistent in his overriding preference for money, and that he is not merely deluded or mistaken about what his true preferences are. But if there is no evidence that Ebenever is mistaken about what he wants, then the economic AI must ultimately conclude that it should help him achieve it.

On the Aristotelian model, by contrast, Ebenezer's end is flawed because it is destructive to his wellbeing. He has the wrong end. His true objective, the one worth pursuing, is the virtue of wealth, not money. Wealth, in this Aristotelian context, is the right amount of money, handled in the right sort of way, compatible with the other ends that constitute a good human life.

So, how should we want our human-centered AI to treat us? Should it only help us to achieve objectives which – through some process – it understands as worthwhile? Or should it help a fully consistent Ebenezer do what he prefers, even if it makes him worse off in some objective sense?

The economic model has appealing aspects. It seems to put each of us in charge of our own destiny. We each maintain a kind of control that seems lost on the Aristotelian view. Ebenezer seems to prefer a life with more than the "right" amount of money. It's odd to suggest that an AI should understand his true ends as something that he manifestly disprefers. If it did, that would seem to be taking Ebenezer out of the center of the AI's assistance task. Relatedly, the economic model seems to reflect egalitarian and democratic ideals. We don't want to prejudge the question of what kind of human life is worthwhile. Each of our views should be treated as authoritative when it comes to actions that only concern ourselves, and equal when it comes to how we live together[14].

It's important to keep in mind the difference between the AI's interpretive and assistive tasks, however. It may be reasonable to claim that Ebenezer should have complete authority over how the AI interprets his objectives. But it's unreasonable to claim that Ebenezer should have complete authority over how the AI assists him in achie-

[14] We discuss these concerns further in the section Doubts about the Aristotelian Approach: Control, Paternalism, and Feasibility.

ving them. Russell himself agrees that we wouldn't want what he calls "loyal" AIs, since they may assist individuals in acts harmful to others[15]. But the issue goes beyond conflicts between individuals in society, since whole societies may prefer bad ends[16]. The Aristotelian AI still would avoid rendering assistance in such cases, but there is no guarantee that the economic AI would, since it determines what is good solely by reference to human preferences.

Moreover, we might want human-centered AI to take some liberties in interpreting our objectives. Most of our deepest concerns are significantly underspecified. We want fulfilling relationships, professional success, and personal growth, but we aren't sure what, exactly, constitutes each of these things. While the economic AI characterizes our objectives in purely descriptive language and takes them to be fully legible in our behavior, the Aristotelian AI assumes our objective is to achieve what is good for us, and it interprets our behavior in light of that assumption.

Incontinence (*akrasia*)

Consider Sherlock the unwilling addict who, despite his better judgment, continues using. The ends he sets for himself are, let's suppose, perfectly acceptable. And he is under no illusions about the fact that use limits his ability to achieve those ends. Importantly, he is not compelled by some overwhelming desire or external force, but voluntarily does what, according to his best judgment, he knows he shouldn't do. There is thus an inconsistency between Sherlock's stated ends and his actions.

[15] S: RUSSELL, *Human compatible: artificial intelligence and the problem of control*, cit., p. 213.

[16] The point applies even in the case of a wrongful global consensus. For example, even if everyone in the world thought slavery was permissible, including slaves themselves, it would still be wrong. For more on this point in a machine ethics context, see A. T. WRIGHT, «A Kantian Course Correction for Machine Ethics», in G. J. ROBSON, J. Y. TSOU, eds., *Technology Ethics: A Philosophical Introduction and Readings*, Routledge, New York, 2023, pp. 141-151.

On the economic model, the question the AI would confront is, What is Sherlock's true preference? The AI might take him at his word and so help him overcome his addiction. Or, the AI might conclude that he in fact prefers to continue using the drug despite what he says, and so would help him come to terms with that. The economic AI must clarify Sherlock's true preferences, before it can decide how to help.

The Aristotelian AI conceives the inconsistency between Sherlock's end and his actions quite differently. It is rooted not in conflicting preferences, but rather in a lack of self-control or weakness (*akrasia*). Sherlock's end is indeed to quit using; he just cannot make himself stop. Akrasia is when the rational part of our soul cedes its rightful place to the desiderative part. When we are akratic, our true objectives are reflected in our deliberation and judgements about what we should do, not in our choices and actions. So the Aristotelian human-centered AI would not take the akratic's actions as indicative of their objectives. It would, instead, pay attention to the deliberation of the akratic agent and help them put reason back in its rightful ruling position.

The fact that the economic AI might conclude that Sherlock prefers to continue taking the drug, despite what he says, seems to be a problem, especially if we assume that his addiction is more or less destructive to his wellbeing. It is difficult to see how an AI that helps Sherlock to continue using benefits him. The case is somewhat like that of Ebenezer, where the economic AI may help him pursue a bad end, except here, Sherlock himself judges it a bad end. Worries about paternalism thus fade, since it would seem more respectful of Sherlock's autonomy to take him at his word. The Aristotelian AI, by contrast, seems to get the case right.

Some have proposed that the economic model might conceive an akratic like Sherlock as someone with a conflict between first- and second-order preferences[17]. His first-order preference – for taking the

[17] Sen suggests that the idea of a second order preference is useful for modeling the akratic (A. K. SEN, *Rational fools: A critique of the behavioral foundations of economic theory*, in «Philosophy & public affairs» 6, 4 (1977), pp. 317-344, p. 340).

drug – is expressed by his desire to take the drug. But his second-order preference – expressed by his judgment that he shouldn't take it – is to not have that first-order preference. He wants to use the drug, but that's not what he wants to want. This proposed innovation seems promising, since the economic AI may then presumably help Sherlock to overcome his first-order preference to continue using, which seems to be the right result.

But it's hard to see the rationale for prioritizing higher- over lower-order preferences on the economic model. Even if akrasia can be modeled as a conflict between first- and second-order preferences, irrationality on the economic model consists only in a kind of inconsistency between them. We can resolve that inconsistency by changing *either* our first- *or* our second-order preference. The economic AI might just as well decide to help Sherlock change his second-order preference to avoid using the drug to better accord with his akratic first-order choice to continue using, rather than the other way around.

Mere continence

Aristotle defines a merely continent person as one who correctly reasons about what is right and chooses to do it but does not enjoy it and gains no pleasure from it. Consider Carnegie, a philanthropist who makes major charitable gifts but does not enjoy it. He gives only because he believes he should, even though it pains him to give away his hard-earned wealth.

Russell considers the related idea of a meta-preference, though he is focused on the problem of preference change, not akrasia (S. RUSSELL, *Human compatible: artificial intelligence and the problem of control*, cit., p. 241). Sen takes second order preferences to express our ethical judgments, but we should note that akrasia doesn't need to be understood as an inconsistency between ethical judgment and action, but between any judgment about what we should do and what we do. On this point, see D. DAVIDSON, «How Is Weakness of the Will Possible?», in D. DAVIDSON, *The Essential Davidson,* Clarendon Press: Oxford University Press, New York 2006, pp. 72-89.

Aristotle understands continent action as not fully rational. On the Aristotelian model, rationality requires a well-formed character, one where the passions and appetites have been trained properly toward the virtues. To act wisely, one must do the right thing at the right time *in the right way, with the right desires and feelings*. One does not merely avoid logical errors in deliberation but also avoids desiderative and affective errors as well. A continent person like Carnegie is aware that his character is not fully virtuous and the AI's role thus might be to provide support mechanisms or interventions to help him overcome this personal struggle.

It isn't clear how to model continence from the economic perspective. The economic perspective may see little or no difference between rational and continent action. If we take the economic model as failing to distinguish between the continent and the rational agent, then it presents a sharp contrast with the Aristotelian perspective. The machine is concerned only with our preferences, not with the pleasures we derive from satisfying them. In this sense, the Aristotelian model recognizes forms of consistency that the economic model leaves out – consistency between our choices, desires and affect.

Perhaps the economic model could invoke second-order preferences again and model the continent person as the inverse of the akratic. The fact that it pains the continent to act as they think they should may indicate a first-order preference against doing the right thing. But the fact that they judge and act correctly indicates a second-order preference for doing the right thing. While the akratic's first-order preference (e.g., to take a drug) wins out in the conflict, the continent person acts according to their second-order preference (e.g., to give to charity).

But just as in the akrasia case, it's not clear why the economic AI would conclude - as the Aristotelian AI would - that it should help the continent person bring their desires and affect in line with their choice. The conflict might just as well be resolved in the other direction.

Wantonness

Consider Emma, a romantic who spends her time fantasizing about fancy parties and passionate love affairs. She thinks of herself as a good person but rarely thinks about what that means. She may give money to someone in need one day when moved by their plight, but then fail to do so the next for someone else in relevantly similar circumstances.

It is possible for humans to act with little or no reflection upon their reasons for action, somewhat in the way of a nonhuman animal. Call someone who acts in such a way "wanton", since her behavior is likely to be less coherent than the behavior of someone who is more reflective. A wanton may sometimes do wrong but other times do right, as the fancy takes her.

How should a human-centered AI assist someone who acts wantonly? From both the economic and Aristotelian perspectives, the best way would seem straightforward: help the wanton reflect on her reasons for action. However, the two approaches have different motivations for assisting the wanton in this way.

From the economic perspective, the observed incoherence in her behavior stems from ignorance or false beliefs about her true preferences, which are assumed to be coherent and stable. If the wanton acts capriciously, that is only because she does not know what she truly wants or how to get it[18]. To assist the wanton, the economic AI may help her reflect in order to discover what kind of life she really prefers and then help her devise a plan to achieve it. The economic AI might also attempt to model the wanton as someone who simply lacks second-order preferences[19]. The AI may thus help Emma develop such develop such second-order preferences by helping her reflecting on her first-order preferences, so that she might then make choices more consistent with her true preferences.

The Aristotelian AI, by contrast, would understand the wanton's erratic behavior to stem from a lack of proper education. The wanton acts capriciously because she has not been taught what ends are good

[18] Note that if the wanton did not behave capriciously, then the economic AI would not distinguish her as a wanton at all.

[19] See H. G. FRANKFURT, *Freedom of the Will and the Concept of a Person*, in «The Journal of Philosophy», 68, 1 (1971), pp. 5-20.

for her and why, or how to properly go about achieving them. She may also lack important habits and dispositions of character requisite for acting virtuously such as temperance. To assist the wanton, the Aristotelian AI may help her to learn the nature of virtue and good reasons for action. It may also help train her character into the virtues.

The purpose of such education and training would not merely be to help Emma behave less erratically, however, as on the economic model, but to act virtuously. Wise action requires not merely doing what is right, but doing it in the right way, at the right time, and *for the right reasons*. Reflection on one's reasons for action is thus necessary to act fully virtuously. Hence even if Emma's behavior were perfectly consistent outwardly with virtuous action, she would still be a wanton on the Aristotelian model.

Doubts about the aristotelian approach: control, paternalism, and feasibility

There are a number of related concerns about the Aristotelian approach to modeling the human in human-centered AI. The fact that an Aristotelian AI would correct for our bad ends seems *paternalistic*: the machine hinders or refuses to assist us in our choices and justifies that by reference to our good. Correcting for our bad ends as well as for our akrasia or wantonness also seems to threaten our *control* over the machine: the machine ignores our voluntary choices and does what it thinks best, instead. And trying to correct for our mere continence seems to make our machines meddlesome. In short, we might worry that the Aristotelian approach takes the human out of the center of human-centered AI.

One response to these concerns is to minimize them. If the AI really did understand the human good better than humans themselves, then perhaps some form of coercion or manipulation would be permissible as a way to help humans achieve it. After all, humans would be forced to do only what they themselves would admit is their good, if they were fully rational. Ebenezer would presumably eventually see the error of his greedy ways after being set on the right path; Sherlock may be grateful for an intervention to stop his drug use; and Emma may reflect with relief that the AI forced her to leave her frivolous life behind. Perhaps we should cede control to meddlesome machines, if that really were for the best.

Another, better, way to respond is to recall that human virtue is *epistemic* as well as practical. Only someone with courage can know exactly what the right thing to do in the face of fear is, and why they should; and only someone with compassion can know how best to help someone in need in some situation, and why. Only humans can have the feelings and affect that serve both as the primary source of evidence about which actions are virtuous and partly constitute virtuous actions. Only humans can have human virtues[20]. The Aristotelian AI would recognize that humans have privileged access to information about virtue and the human good. Of course, not every human will have superior insight into the human good, and the machine should not take the behavior or beliefs of obviously vicious or akratic or wanton humans to indicate what is good for them. But a machine attempting to learn what virtue is by observing us could never just ignore what a human does or believes, even one that often acts foolishly. Humans as they are would therefore remain at the center of the Aristotelian approach to human-centered AI.

This response helps address the concern with control, but not the concern with paternalism, at least not fully. There are two ways to frame this concern. The first is to assert that only humans should have the right to determine what is good for them and what they should do, even when they may be wrong on both counts. If Ebenezer thinks greed is good, then perhaps it is his right to think so, even if he is wrong, and no machine should have the authority to question his obsessive pursuit of riches, so long as he harms no one else. For that

[20] Only humans can have human virtues because, according to Aristotle, virtues are states that make a thing perform its characteristic "function" (ergon) well, and by doing so, realize its good (ARISTOTLE, *Nicomachean Ethics*, cit., 1106a1-2). Human virtues are thus those states that make humans perform the human function well, which under good conditions constitutes the human good.

matter, an AI should respect even akratic choices, so long as they are voluntary. Any coercive interference with Sherlock's freedom of choice can be justified only by his *actual*, not merely hypothetical, consent. It is not enough to say that someone *would or should* have consented to some interference with their rights, when they didn't.

The second, related way to frame the concern with paternalism is rooted in the modern democratic commitment to reasonable pluralism. Any perfectionist argument to establish a conception of what is good for us will be subject to reasonable disagreement in a modern pluralistic society like our own. Citizens' commitment to reasonable pluralism rests on their recognition of what John Rawls refers to as the "burdens of judgment"—factors that make it difficult to determine correct answers to questions about the human good and what one should do. Reasonable citizens are aware of the burdens of judgment and are, therefore, unwilling to impose their own comprehensive conceptions of the good upon those with different conceptions[21]. But that is just what the Aristotelian AI may attempt to do. It would judge Ebenezer's pursuit of money, for example, as incompatible with its comprehensive Aristotelian account of the human good and so would hinder or avoid assisting him in its pursuit.

The simplest response to these concerns with paternalism may be to just reiterate that we do not want an AI to assist humans in the pursuit of harmful ends. The human-centered AI, therefore, must form some conception of the human good, so that it can avoid assisting us in doing what is harmful. And that is the core of what the Aristotelian approach to human-centered AI recommends.

But this simple response seems insufficient because there are alternative, thinner conceptions of the human good more compatible with rights and pluralism than is any robust form of Aristotelianism. Perhaps Mill's utilitarian liberalism would suffice. We would then hold that a beneficial AI should defer to human decisions about what is good except when those decisions would cause harm to others,

[21] J. RAWLS, *Political Liberalism*, Columbia University Press, New York 1993, pp. 36-37.

where harm is defined in terms of damage to interests protected by rights that touch on the "essentials of human wellbeing[22]." The point is that a thinner account of the human good may suffice to vindicate the intuition that an AI should not assist us in harmful acts, while better respecting our rights to make our own choices about what we should do. These thin accounts of the good would not be fully compatible with Rawlsian political liberalism, either, but they seem to raise less of a concern with paternalism than does the Aristotelian account. A thin (neo-) Aristotelianism account may suffice as well, but such an account may differ little from a Millian or even Kantian liberalism[23].

A final concern with the Aristotelian approach to human-centered AI is with its feasibility, keeping in mind that AGI of any kind is still quite speculative. The AI that adopts an economic model of our rationality needs only to observe our choices to try to discover our revealed preferences. The AI that adopts the Aristotelian approach needs to consider, not just our choices, but all the relevant facts about how the objects of our choices might contribute to the realization of objectively good human ends[24]. Not only are there more inputs for the Aristotelian AI's learning problem, but the inference from observed

[22] J. S. MILL, «Utilitarianism», in J. M. ROBSON, ed., *Collected Works of John Stuart Mill*, vol. 10, p. 255.

[23] Another possible response is to concede that the Aristotelian approach is incompatible with a modern pluralistic society (see A. MACINTYRE, *After Virtue: A Study in Moral Theory*, University of Notre Dame Press, Notre Dame 1981). But if so, then the concern with paternalistic AI seems even more serious.

[24] Note than a *non*-human-centered Aristotelian AI avoids this complexity. An AI that is not human-centered would be programmed to learn a model of wise human action by observing exemplars of virtue identified as such by its programmers, and then it would help or hinder humans to act in accord with what it thought one of these exemplars would do in their place. It need not evaluate whether the exemplars were actually acting in ways that realize the human good. The risk, however, is that the AI's model of virtuous action would be incorrect or incomplete, perhaps because its programmers specified too few or the wrong exemplars. This is a version of the specification problem.

human behavior to objectively good human ends seems less secure than the inference from behavior to preference.

But while the learning problem for the Aristotelian AI seems more challenging, that isn't necessarily an argument against it. The appeal of the economic model stems, in no small part, from its formal rigor. We have not yet developed the formal tools needed to make the Aristotelian model suitable for applications in human-centered AI. If we did, we might find that the learning problem, though difficult, is still manageable.

Conclusion

We've suggested that elements of the Aristotelian model of human rational agency are important for designing truly beneficial human-centered AI. The economic model was developed to predict human behavior in the marketplace. In that context, the assumption that we generally behave in accord with our preferences and work to maximize our utility makes sense. However, human-centered AI is not trying to predict our behavior, but to assist us in achieving our objectives, not just in the marketplace, but across a broad range of our activities. In that context, we think more substantive theories of rational action can provide a better framework for structuring the AI's problem of learning our objectives from our behavior.

Abstract

As we build increasingly intelligent machines, we confront difficult questions about how to specify their objectives. One approach, which we call human-centered, tasks the machine with the objective of learning and satisfying human objectives by observing our behavior. This paper considers how human-centered AI should conceive the humans it is trying to help. We argue that an Aristotelian model of human agency has certain advantages over the currently dominant theory drawn from economics.