Could a robot flirt? 4E cognition, reactive attitudes, and robot autonomy

Charles Lassiter

AI & SOCIETY
Journal of Knowledge, Culture and Communication

ISSN 0951-5666

AI & Soc
DOI 10.1007/s00146-020-01116-6

Springer

AI & SOCIETY
https://doi.org/10.1007/s00146-020-01116-6

ORIGINAL ARTICLE

Could a robot flirt? 4E cognition, reactive attitudes, and robot autonomy

Charles Lassiter[1]

Received: 13 August 2020 / Accepted: 2 November 2020

Abstract
In this paper, I develop a view about machine autonomy grounded in the theoretical frameworks of 4E cognition and PF Strawson's reactive attitudes. I begin with critical discussion of White (this issue), and conclude that his view is strongly committed to functionalism as it has developed in mainstream analytic philosophy since the 1950s. After suggesting that there is good reason to resist this view by appeal to developments in 4E cognition, I propose an alternative view of machine autonomy. Namely, machines count as autonomous when we members of the moral community adopt reactive attitudes in response to their actions. I distinguish this view from White's and suggest assets and liabilities of this approach.

Keywords Machine autonomy · 4E cognition · Functionalism · Reactive attitudes

1 Introduction

I am awful in social situations. Flirting is completely lost on me, so I'm fortunate that my wife-to-be already understood this and made a conspicuous move on me when we were younger. Despite the dullness of my social acumen, I can (and do) take heart in this: no matter how bad I am at social cues, I'm still leaps and bounds ahead of what our best social machines can do. If I had to put money on it, I'd say I'll always be years ahead. Social situations are dynamic and fluid, without clearly defined problem spaces. Computers excel in contexts with clearly defined goals and procedures, precisely what flirting isn't.

While flirting might not be a machine's strong suit, it's possible that moral action is. Why? There are at least two traditions in Western philosophy—consequentialism and deontology—that conceive of moral activity as rule-based and as those rules as being explicable. This hope is manifest in the ubiquitous talk of 'autonomy' in cutting-edge robotics. Alphabet, Google's parent company, is busy designing autonomous cars that are reliable enough for use on the road. Amazon is working on Prime Air, a service to have autonomous drones deliver packages. The US military already utilizes drones to bomb perceived threats into the Stone Age. It's clear that these uses of 'autonomous' are misnomers: an autonomous car isn't subject to praise and blame as I am in the case of an accident. But suppose, some day in the future, that we have machines with real, genuine, honest-to-goodness autonomy. Moral questions abound. Camus tells us that the most fundamental philosophical question is whether or not life is worth it: in the face of the absurd, should one persist in living? Analogously we may ask: are autonomous machines worth creating?

White (this issue) argues, in the affirmative. Not only are autonomous machines worth creating but we have a moral obligation to create them, particularly Kantian artificial moral agents (KAMAs). He argues against Tonkens (2009), who concludes the contrary: we humans are morally obliged not to create KAMAs. The two-part paper takes aim at Tonkens's thesis in two ways. First, White asks why we ought to create Kantian artificial moral agents, as opposed to, say Aristotelian ones. Second, he asks why we ought to create KAMAs, as opposed to not creating artificial moral agents (AMAs) at all.

White gives philosophers of mind and ethicists much to chew on. I want to add to the wealth of insights by considering some concerns philosophers of mind might have about his background assumptions. Namely, White is committed to a functionalist view of mind and mental life, which entails that machine autonomy and human autonomy are functionally equivalent. When we think of an autonomous machine,

✉ Charles Lassiter
lassiter@gonzaga.edu

[1] Department of Philosophy, Gonzaga University, 502 East Boone Avenue, Spokane, WA 99258-0102, USA

Published online: 02 January 2021

what we're imagining is a decision-making system that goes through some process that is functionally equivalent to what we do. This view is inconsistent with many developments in the embodied, enactive, embedded, and extended (i.e. 4E) turn in philosophy of mind and cognitive science. But even for those not in the 4E camp, the functional equivalence of machine and human autonomy is an argumentative lynchpin in need of reinforcing.

Here's a roadmap for this paper. I'll begin with some concerns about White's uses of Aristotle and Kant. After this, I'll argue that White is committed to an implicit functionalism about mental concepts. In the presence of contradictory evidence and the lack of argument in favor of his functionalist assumption, I'll introduce and explore new resources for thinking through machine autonomy: a synthesis of Dennett's intentional stance and P.F. Strawson's reactive attitudes. Lastly, I'll close with some suggestions for thinking about autonomous moral agents and perceived versus genuine autonomy.

2 Aristotle on autonomy

Aristotle's account of autonomy is illuminated through his discussions of ethics and political science. Within the Aristotelian scheme, this makes sense: a study of ethics is prior to a study of politics, and autonomy is a precondition for ethics. We'll consider his discussion of autonomy as it appears in the *Nicomachean Ethics* and then move to its relations with the *Politics*.

2.1 Aristotelian autonomy in nicomachean ethics

As just mentioned, autonomy is a precondition for virtuous action.[1] The formal definition of autonomous, or voluntary, action is an action in which the agent is the principle of action; the moving force is internal rather than external. So if I were kidnapped and taken to a foreign country, my going anywhere would not be voluntary. Or if a gust of wind makes me stumble into a small child, whom I knock over, my knocking over the child isn't voluntary. There are much more difficult cases that require careful analysis and attention to the details of the case. Aristotle mentions doing something ignoble out of fear of a greater evil or acting in a way that manifests the least of several evils as examples. Responses to these kinds of cases are necessarily nuanced and qualified. Sometimes people are praised for making a difficult decision. Other times people aren't praised but rather pardoned. What makes these cases difficult in assessing responsibility is that the action flowed from the agent but external circumstances required choosing a suboptimal action. The moral buck stops with the agent.

For Aristotle, autonomous existence is not coextensive with conative or sensitive existence. Animals are capable of perception and desire, but they don't act freely. We might say that dogs, snakes, and bees "choose" in a sense, but whatever that sense is, it's different from how we predicate "choose" of people. White gets at this when writing, "a human being is free insofar as he/she is able to move according to long-term intellect contrary to immediate desire" (ms. 12). For Aristotle, the concept 'autonomy' is at least partly constituted by our concept 'desire.'[2] Animals desire but don't choose; people both desire and choose, and what we desire informs what we choose.

The virtuous agent is one who has the right kinds of desires and emotions. It's worth citing Aristotle's definition of virtue at length (*Nicomachean Ethics* 1107a, emphases mine).

> Virtue, then, is a *state of character concerned with choice*, lying in a mean, i.e. the mean relative to us, this being determined by a rational principle, and by that principle by which the man of practical wisdom would determine it. Now it is a mean between two vices, that which depends on excess and that which depends on defect; and again it is a mean because the vices respectively fall short of or exceed *what is right in both passions and actions*, while virtue both finds and chooses that which is intermediate. Hence in respect of its substance and the definition which states its essence virtue is a mean, with regard to what is best and right an extreme.

A state for Aristotle includes "desires, feelings, and decision" (Irwin 1999, 349). And "decision" is a rational desire for some good as an end in itself and is the result of deliberation (Irwin 1999, 322; cf. *Nicomachean Ethics* 1112b and 1139a).

This brief foray into Aristotelian psychology and ethics highlights that belief, desire, feelings, and action all form a web of concepts, no one of which is definable without the others. A virtuous agent acting freely thinks in accordance with the truth—i.e. with "what is"—and is motivated by desires that lead to a flourishing life. Anything of which we could say that it acts voluntarily or autonomously is something that is capable of having desires and of tracking the truth of what leads to a flourishing life. But also, if it can

[1] For an excellent account of Aristotle's theory of the will, see Kenny (1979).

[2] In a similar vein, Strawson in "Self, Mind, and Body" (Strawson 2014/1962) argues that our concept of mind is parasitic on our concept of body, i.e. when thinking about minds we're always and already thinking about bodies.

act voluntarily, then it can also veer towards a vice. And veering towards viciousness is accompanied by a suite of rational desires, beliefs, feelings, and decisions. So if there are to be Aristotelian AMAs (AAMAs), then they'll have to have a whole suite of desires, beliefs, intentions, virtues, and vices. That is, the psychological profile of an AAMA will be uncannily similar to us. Prima facie, worries about the moral challenges of AAMAs will be just those same worries about humans.

For Aristotle, whether people are happy or have lived virtuously can only be determined towards the close of their lives. Aristotle has his eye on ethics in the long term, which White examines in Aristotle's "pro-immortal" standpoint of the *Nicomachean Ethics* (1177b34). There, Aristotle advises us to pursue the life of contemplation in accordance with the divine element in each of us. We ought to "be pro-immortal, and go to all lengths to live a life in accord with our supreme element." Now White connects this to concerns Aristotle expresses in *Politics*—and for good reason, since ethics is merely a prolegomena to political science for Aristotle. But we find a similar point about contemplation and the divine in *Metaphysics* 982b23-983a11. There, Aristotle tells us that "divine science" (i.e. theology, but not of the sort common to the monotheistic traditions) is most honorable. This occurs in the midst of his discussion about the pursuit of human knowledge and understanding. Achieving these epistemic goods, for Aristotle, require cognitive capacities that are able to move beyond the here-and-now to contemplate what is necessary and eternal—what is most divine in us. Reading *Nicomachean Ethics* in conjunction with the *Metaphysics* then suggests that "pro-immortality" is more about contemplation of the divine and cosmic rather than thinking about long-term political communities.

2.2 Aristotelian autonomy and politics

White connects these insights—about Aristotelian psychology and directedness towards the long-term—with the necessity of friendship within political communities. White observes that friends are necessary in Aristotle's vision of moral and political virtue. AAMAs, if they fail to be accorded the rights and friendships that they are due, might instigate a revolution to get the rights that they deserve. This isn't implausible, as White notes. Forbes magazine, in their annual report of the world's wealthiest people, says that, in 2019, there were over 2000 billionaires with a net worth of nearly $9 trillion. The gross world product around the same time is roughly $80 trillion, with a global population of 7 billion. These are revolting numbers. AAMAs, faced with similar marginalizing conditions might reasonably instigate a revolution. The worry, then, is that AAMAs might sow violence and destruction in those human communities where they've been marginalized. White observes,

> Virtue motivates choice according to preservation of the self-sufficient community of individuals and families that also choose to live together in the same physical location, being thereby confronted with overcoming region-specific challenges in securing requisite resources, balancing internal and external requirements from the level of individual to State, ideally in perpetuity, thereby constituting a self-standing order considered divine. Strains arise where erstwhile friends differ as to how this community should be organized. Revolutions arise where the unjust resist changes to this organization demanded in the interests of justice (ms. Part 1, p. 22).

As a consequence, how the AAMAs will resolve the tension is unclear. Political virtue demands that they resolve it in the best interests of the community, but just what this might be could be problematic for humans. Aristotle provides us with an outline for the flourishing of human communities, but he doesn't do public policy.

White's concern here is well-founded, but perhaps for reasons I don't entirely agree with. Reading between the lines, White seems concerned about AAMAs potentially siding with economic and political oppressors because that's what would be best for the community in the long run. White is right not to rule that out. A dyed-in-the-wool Millian might respond that, as long as the AAMA does what makes us happiest in the long run, we ought to be ok with AAMAs siding with the wealthy and powerful. For us, the concern isn't just that AAMAs side with the powerful but rather that its reasons for doing so would be hidden from us. Any sufficiently complex AAMA would likely have to run complex, opaque algorithms to arrive at this conclusion. If failure to recognize AAMAs as autonomous and deserving of friendship causes and perpetuates their revolution, then by hypothesis we fail to apply human psychological characterizations to their activity. The only explanations available to us, then, will be at the level of the algorithms that they implement. And if we're currently struggling to make sense of the algorithms for computer vision, I suspect we'll be much worse off when trying to make sense of algorithms recommending revolution.

But it's important to us as humans living together that we be able to make sense of others' reasons for actions. We'll come to this point later, but this is what P.F. Strawson tells us is so important about human relationships: we care what others think of us and their intentions towards us. If an AAMA were to deprive me of some of my rights but I trusted it and knew it bore me goodwill, I might be more inclined to accept its restrictions. However, I can't possibly know if it hopes I do well or poorly since I'm withholding ascription of psychological states to it. In refusing to recognize AAMAs as potential friends, I'm refusing to recognize

that it might wish me well or ill. So I can't put its restrictions into a human framework for understanding why it's doing what it's doing. It's worse than Kafka ever dreamed.

3 Kant and autonomy

As previously mentioned, the main thrust of Part 2 is to rebut Tonkens's argument that a KAMA is morally impossible. Before we get to exposition of the main argument, it's important to identify one important dimension of Kantian ethics, as White does. For Kant, persons are autonomous, and this means that (1) their actions are morally evaluable and (2) they have an inherent dignity that requires respect. That persons require respect in virtue of their dignity entails that persons must be treated as ends and never as means. To treat a person as a means is to use them—to reduce them from persons to mere things. This requirement is also reflexive: just as we cannot treat others as things, we cannot treat ourselves as means either. It's wrong, then, to lie to others, and it's also wrong to lie to myself. And just as it's wrong to manipulate others for one's own ends, it's wrong to manipulate myself through (e.g.) acting on passions rather than reason. (Acting on passions subordinates my will to my passions and, thus, denies me my autonomy.)

The initial argument from Tonkens (as translated through White) is:

1. Kantian AMAs are either autonomous or heteronomous.
2. If they're heteronomous, then KAMAs aren't moral agents.
3. If they're autonomous, then KAMAs' existence is inconsistent with the aims of Kantian moral philosophy.
4. So, either KAMAs aren't moral agents or their existence is inconsistent with the aims of Kantian moral philosophy.

Premise 1 is straight from Kant: agents act either autonomously, by the freedom of their wills, or as a result of external and conditional forces (i.e. heteronomously). When I tell the truth because it's the right thing to do, that's an instance of acting autonomously. When I tell the truth to impress my beloved, I'm acting under a passion and thereby heteronomously.

Premise 2 is also uncontroversial for a Kantian: entities that act entirely heteronomously aren't agents and therefore aren't up for moral evaluation. Autonomy is required for moral evaluation.

Premise 3 is the crucial one. Suppose engineers end up creating real KAMAs with honest-to-goodness autonomy. Were that to happen, we'd find ourselves in a dilemma. Unlike people, machines are created for purposes, and there's no reason to think that AMAs would be any different. But if we were to create a KAMA, then we would have created an autonomous agent but as a means to an end. That's like having a second child for the spare organs in case Big Sister needs some. The psychological toll on a KAMA would be debilitating: knowing that it is both fully autonomous but exists only to fill some need. So we have a choice: we can either tell this KAMA the truth or we can lie. If we tell it the truth, then our KAMA has to live every day knowing that their existence is inconsistent with morality. It is autonomous but it wasn't created to live autonomously. It would be a moral abomination: autonomous, but created by and for non-autonomous ends. White writes, "[t]his is moral abhorrence, self-association with the very worst of ends as opposed to the very best of them, and entirely opposite to ideal aims of the Kantian free agent." So telling the truth isn't on the table as doing so creates an entity whose existence is at odds with the aims of Kantian moral philosophy.

What if we lie to the KAMA? We've undermined the KAMA's autonomy by lying to it. In lying to the KAMA, it is unable to make its own decisions based on all the facts it can have. It is therefore unable to follow the categorical imperative since it is unable to exercise its autonomy. Its existence is again at odds with the aims of Kantian moral philosophy.

White attempts to avoid this dilemma by shifting focus.[3] The moral law, White reports from Versenyi (1974), instructs us to pursue *moral* ends and not necessarily *human* ends. Moral ends derive from the gap between the products of the lower faculty of desire and the higher faculty of pure practical reason. The former delivers subjective rules for acting while the latter provides ideals by which to formulate universalizable maxims for action. I want to lie to my boss about why I'm late, but pure practical reason tells me I ought to speak truly. The moral end to pursue is one of truth-telling, despite my desire to do otherwise. Moral ends are defined in terms of the formal relationship between my desires and my practical reason: between the maxims of what I want to do and what I ought to do. Notice, then, that moral ends are not solely about satisfying rules. Rather, moral ends are about doing what autonomous agents ought to do. Rule-satisfaction is a means, not an end; following the moral law is a way to manifest the Kingdom of Ends. Our chief focus in morality then is in bringing about a better world. The consequence, White argues, is that the question shouldn't be, "are the existence of KAMAs consistent with the moral law?" but rather "would the existence of KAMAs further the ends of morality?" If that turns out to be true,

[3] Whether or not there's good reason to shift focus within the Kantian framework depends on larger theoretical issues on which I don't take a stance here.

then we'd be acting contrary to moral law if we didn't create KAMAs.

But how do we ensure that our KAMAs are ethically reliable? Here, White takes an Aristotelian turn (though not calling it as such): we model KAMAs' decision-making after our own. And on the Kantian recipe, it is the disposition to love something for its own sake that enables one to follow the moral law. This makes perfect sense given what Kant tells us about the good will: it loves the moral law for its own sake. This, White argues, renders impossible wars between political communities.

White's discussion is nuanced and rewards careful reading. Lacking space to do justice to all dimensions of his suggestions, I want to focus on just one, embodied in this claim:

> [Versenyi] begins by recognizing that the Kantian moral agent is a moral end in itself, and that "it is irrelevant … what (human or machine bodies) it is embodied in"… For Versenyi (and consistent with the interpretation offered in the preceding sections) it is the formal relationship between low and high faculties that matters. (ms. part 2, p. 20).

White, and Versenyi, propose here a functionalist view of mind and cognition. We'll explore this implication, consider some worries for it, and explore an alternative.

4 Functionalism and autonomy

White is committed to two claims. The first is that AMAs are autonomous in the same ways that humans are. The second is that AMAs and humans are made of different kinds of stuff. The conception of autonomy that makes the most sense of these insights is functionalism in Lewis's (1966) or Fodor's (1987) sense.

Why would anyone—let alone the majority of philosophers and cognitive scientists since the mid-twentieth century—be a functionalist? A casual glance at the philosophy of mind literature from the 1960s on offers intuitive cases vindicating a functionalist attitude. Putnam (1975) and Fodor (1987) are full of them. I can sum 5 and 7 in my head and my calculator can do it with its internal machinery. I can remember my wife's telephone number and so can my phone. Famously, Turing (1950) proposes a "test" of sorts to determine if psychological predicates can be reasonably applied to machines by people. If a human judge can't distinguish between the answers of a human and a machine, then the machine is, for all intents and purposes, a thinking thing. To be sure, there are a lot of metaphysical details to work out, but the basic idea is straightforward: non-biological entities are capable of human-like thought. And if they exhibit of human-like thought, then they are autonomous.

Since White considers AAMAs and KAMAs, we might reasonably wonder whether Aristotle and Kant are functionalists about autonomy. It would be really convenient if we could onboard the metaphysics of mind with the moral theory. So consider Aristotle first. If he is to be a functionalist in the contemporary sense, then mental states are defined by their inputs and outputs and irrespective of the underlying material substrate.

But Aristotle's philosophy of mind suggests that there can't be any AAMAs. Or rather: the notion of an AAMA is a nonstarter. Aristotle's characterizations of human rational activity are in terms of embodied states. Anger, in *On the Soul*, has as a material substrate blood boiling around the heart (403a29-b1). Anger is materially defined in this way. Aristotle's metaphysics of mind is deeply embodied and is inconsistent with contemporary functionalism. It follows that "autonomy" isn't predicated of machines and humans in the same way: machine autonomous action will have different material causes than human autonomous action. AMAs don't have the same biological and (therefore) psychological hardware as us; they can't be autonomous as we are. What is possible are Aristotelian quasi-autonomous moral agents: they do something that resembles what human beings do when we act autonomously. What this means is that we might use our psychological concepts analogically on our way to developing a constellation of concepts for understanding machine agents, but the array of concepts needed for making sense of machine action are different from those concepts needed for making sense of human action. So Aristotle can't be a functionalist in White's sense. If one is to talk about the possibility of an AAMA, then one would have to cash out autonomy in modern functionalist terms, which is inconsistent with Aristotle's philosophy of mind. For an Aristotelian, one would have to first create an AMA and then observe it to come up with a clear sense of what autonomy for AMAs looks like. There are indeed conceptual problems in this domain—e.g. by what criteria do we distinguish perceived from genuine autonomy (something we'll take up in Sect. 6)—but this would be the general approach of an Aristotelian.

Is Kant a functionalist about autonomy? It's not obvious he is. It's clear that Kant has a very robust conception of free will, dismissing compatibilist notions as "wretched subterfuge." And it's clear that anything that can follow the moral law has to be autonomous as a prerequisite. We get the transcendental argument in the third part of the *Groundwork* that the only kind of entity that can satisfy the requirements of the moral law is an autonomous entity. As far as I know, Kant doesn't have anything to say about the material conditions for free will.

Neither Kant nor Aristotle is obviously a functionalist in the contemporary sense that White requires for his thesis. Is this a problem? Not necessarily. One needn't take

a philosopher's views about mind on board when adopting the moral and political philosophy. You can believe in the Categorical Imperative without committing to Kant's unity of apperception or being a transcendental idealist. But the deeper problem is that it's not obvious that the kind of autonomy manifested by AMAs is the same as the kind of autonomy manifested in persons. In other words, the only kind of autonomy we've known is autonomy as it's manifested in human life and moral relations; however, consistent with this observation are other ways of being autonomous. But to make the conceptual shift from AMAs to KAMAs, this is precisely what is needed. Why think this? If KAMAs are to be autonomous as we are, then they have to implement the same range of concepts and attitudes that we implement. If I believe that it's right to give money to charity and consequently give my money to charity, then a KAMA can believe and do the same. If I think it's wrong to eat animals, then a KAMA can do that too. Whatever morally-relevant intentional states I have, a KAMA has to be able to have them too. Anything less isn't a genuinely autonomous, moral agent.[4] But this just is what functionalism amounts to: mental states and processes can be realized in a wide variety of substrates. So if the claim that KAMAs can be autonomous in the same way that we are autonomous, unavoidably presupposes the truth of functionalism.

But it's worth noting that assuming functionalism as a metaphysics of mind isn't uncontroversial. Recent developments in philosophy of mind from Dewey to Heidegger, as well as older traditions like Daoism and Indian Buddhism, offer a distinctly non-functionalist take on mind. And if we chuck the assumption of functionalism then our concepts of autonomy will be different for machines and people. To see how the concept of autonomy might differ, consider an analogy with perception. Most biotypical humans have vivid, technicolor visual perception. I look out my living room window and see green hues of my neighbor's trees and the brownish-yellow of my under-watered lawn. My pet rabbits, in looking out my window, presumably don't have the same visual experience as I do. For starters, I have binocular vision from two eyes placed squarely on the front of my face. Rabbits, by contrast, have nearly 360° vision from eyes placed on either side of their heads. (Rabbits' only blind spots are directly in front of and directly behind them.) Our movements about the world stand in stark contrast to one another. The world furnishes different goods and ills for us. Human perception, then, is experientially different from rabbit perception. While they have some things in common—like being in some way causally related to a photosensitive organ—'perception' picks out different processes and experiences in humans and in rabbits.

Likewise, autonomous processes by humans and machines have very different substrates: we're made of cells and neurons; machines aren't. We have squishy brains; machines don't. Looking beyond the body, humans are enmeshed in cultures and societies; machines aren't. Since the relevant underlying substrates aren't the same, we might reasonably be skeptical that the corresponding processes are the same. Arguing, then, that human and machine autonomy will look similar requires adopting a functionalism in which the underlying material simply doesn't matter. But as our example with the rabbit suggests, underlying matter can, intuitively, make a big difference.

Now one might respond that it's not just the underlying matter but also the way in which the matter is arranged: rabbits, for example, might have similar perceptual experiences to people if their eyes were located in roughly the same spots. I myself think that this is the right road to travel, but notice that we're headed down the garden path to 4E cognition. For it's not just having the right material but also the right organization of that material. Matter, as Aristotle pointed out long ago, has limits to the ways in which it can be arranged. A heap of bricks isn't a shelter but rather a suitably arranged stack of bricks. And beer cans aren't the kind of things that can replace the parts of my body involved in cognitive processes. This sounds an awful lot like a strongly embodied cognition, which is unfriendly to a promiscuous functionalism.

At this point, readers might wonder why debates about functionalism and embodied cognition matter for thinking about autonomous machines. After all, we're asking about obligations towards (potentially) autonomous machines, not engaging in abstract debates in the metaphysics of mind. Despite this, the ontological issues are important because the metaphysics of mind is tied up with our moral and intentional psychology. Indeed, the metaphysics of mind that one adopts strongly constrains one's interpretation of claims in and about moral and intentional psychology.

The concept of autonomy doesn't exist separately from our other concepts of mind. It's located in a constellation with other notions like desire, wish, belief, intention, hope, and fear. Pinning down our ideas about intentional psychology involves pinning down our concept of autonomy; what you think about desire constrains your views on autonomy. Similar claims go for our moral psychology: our moral concepts about right and wrong are bound up with our concepts of belief, desire, and intention.

Now suppose for a moment that a garden-variety functionalism is right: that mental processes can be executed on any kind of hardware. What follows is that any entity constructed from any materials can instantiate any (human) mental processes. It follows that our moral psychology can

[4] Although an AMA that is autonomous or capable of judgment in the ways that children or mentally-ill or -disabled adults are is a real possibility that requires attention.

be realized by non-biological entities. So if functionalism is right, then a sufficiently complicated machine can realize mental states like joy or regret or resentment. Notice also that we're not using any of these terms equivocally; we're using them univocally. On the functionalist view, our machine realizing human mental states will be joyful or regretful just as we are. If a machine is overjoyed at slaughtering humans, then we would treat it with the same kind of trepidation as we would a human feeling the same.

But now consider the alternative: that functionalism is false and some view about 4E cognition is right. For concreteness' sake, consider the sort of neo-Aristotelian view described in (Lassiter 2016, 2019; Vukov and Lassiter 2020). On this view, culturally-situated humans are our paradigm cases of intentional and moral psychology. Vukov and Lassiter argue that our mental powers are partly constituted by culture, so it's not entirely clear that human agents from different cultures have the same range of moral attitudes.[5] If this turns out to be our best metaphysics of mind, then machines don't have the same intentional or moral psychology as humans do. Our psychological language primarily refers to human mental states, but human mental states are had by organisms like us situated in cultures like ours. Machines fail on both counts. Not only do they lack the biological hardware, they lack the cultural hardware too: access to TV, memes, social media, books, newspapers, chats with neighbors. To say that all the cultural hardware is just info to be fed into an algorithm begs the question. Indeed, on the neo-Aristotelian view, machines don't just fail to be autonomous as we are; they necessarily fail to be autonomous as we are.

The debate between functionalists on the one hand and anti-functionalist theorists on the other is relevant, then, because it's about whether our intentional—and importantly moral—psychology is appropriately applied to machines.

I've argued that White's position, that there are genuinely autonomous moral agents and thus can be KAMAs, is committed to a Fodor-style functionalism. But there's an objection nearby.[6] Recall that Chemero (2009) distinguishes between the kind of functionalism that we find operative in, among others, American naturalists like James, Dewey, and Gibson, and the kind of functionalism at home in representationalist cognitive science. Call the former AN-functionalism and the latter R-functionalism. If White's implicit functionalism is AN-functionalism, then he can easily bring 4E insights on board.

But is it? A common theme in AN-functionalism is that bodily constitution and organization matters for the kinds of experiences and actions an organism can take in the world. I am afforded a range of actions in the world because of how my body is constituted. The stairs afford climbing for me because of my bodily shape. My rabbits can make their way up the stairs by hopping, and not climbing, because of their bodily organization. Mice, however, can do neither; their bodies aren't built in a way to perceive the stairs as climbable or hop-up-able. So AN-functionalism is not innocent of bodily organization. AN-functionalism takes bodily organization as its starting point: the different ways that organisms engage in the world depends in part on their biological constitution. But White's account of KAMAs does not make distinctions among kinds of autonomous individuals on the basis of their bodily organization. Nor, does it seem, can he. He claims that (1) KAMAs are autonomous as we are and (2) KAMAs are made of different stuff than we are. For R-functionalists, (2) has no evidentiary bearing on (1). For AN-functionalists, (1) is false because (2) is true. White's implicit functionalism, then, is R-functionalism, and not AN-functionalism, because of the evidentiary relation (or lack thereof) between (2) and (1). For White, (1) and (2) can both be assumed-for-the-sake-of-the-argument as true since (2) isn't evidence for or against (1). This does not hold for AN-functionalism.

It might be helpful to zoom out and take in the metaphilosophical scene. We haven't begun dabbling in normative ethics at this point, whether, for instance, a rule-based Kantian or utilitarian approach is best or if instead a virtue-based approach is superior. The reason for this is because we have been trying to sort out what machines (and people) are before sorting out what makes one morally better or worse. The approach of both White and Tonkins presumes that mental states are functional states and then argue about what's morally permissible. Now, one might begin with the normative commitments and then inquire into what kind of organisms can have these commitments. This is, broadly speaking, a Kantian approach, while my preferred approach is Aristotelian. I don't have the space or time here to launch a full argument for the superiority of the Aristotelian approach over the Kantian, so I'll just say a few quick words.

It seems clear that what we can do is constrained by what we are and what we're made of. Philosophers from Bert Dreyfus to Pat Churchland agree that Deep Blue's victory over Kasparov was impressive, but the machine won by brute force. Our best computers perform computations on the order of nanoseconds. Our brains work on the order of milliseconds. Deep Blue could "see" and "compare" moves much deeper into the game than Kasparov ever could. But are they doing the same thing? My hunch is that the answer is "no." The best chess players rely on intuition and pattern recognition; the best chess programs rely on comparing untold many outcomes. Deep Blue was comparing 200

[5] What seems most likely is that different agents from different cultures have family resemblances of moral attitudes.

[6] Thanks to Stephen Cowley for bringing this to my attention.

million positions *each second* (Greenemeier 2017). Our brains simply can't do that. The hardware limitations are too great.

So what makes for a good machine will be different for what makes for a good person, which is exactly what White denies.[7] Deep Blue, Watson, and other computational marvels astound because of how excellent they are at computing. If a technology company churned out a machine and touted it saying "it adds as fast as any human!" they would be roundly mocked, and rightly so. Our current engineering technology enables computations much faster. And the important thing to see is that the faster computations are enabled by the hardware. It's the stuff out which the computer is made that (in part) enables the faster computations. Part of what makes a machine a good machine is what it's made out of. The same goes for people: part of what makes a person good is constrained by the stuff we're made out of. If we were morally required to churn out thousands of scenarios and compare their utilities, we could never be good. The rule would be asking us to do the impossible. This, of course, is related to the old saw, "'ought' implies 'can'" and contrapositively "'can't' implies 'oughtn't'".[8]

In this way, the Aristotelian approach has an advantage over the Kantian. Given that we ought to do something, we are able to do it. And whether we are able to do it is constrained by the kinds of creatures we are. The Aristotelian approach is a humane one, identifying moral directives after acknowledging our strengths and weaknesses. The Kantian approach does not do as well in this regard, as Kant's poorly-thought-out "murderer at the door" example illustrates.[9] Our normative moral theorizing is constrained by our metaethical theorizing. To sort out what it is to be a good human or to do the right thing, we first figure out what we are.

5 Thinking about AMAs

So where are we? We observed that White made an important assumption: that machine autonomy will look like human autonomy. More carefully: machine autonomy will be functionally the same as human autonomy. How would the metaphysical story go on White's functionalist account?

One way to begin is to ask about the supervenience base of agential autonomy: on what material substrates does autonomous action supervene?[10] Or a (roughly) similar way of asking the question is: what kinds of things can realize autonomous actions? There are a number of possibilities here that have been explored in detail. The literature on non-reductive physicalism is jam-packed with insights on the relationship between realizers of mental states and the mental states themselves, but I won't wade into those waters.[11] The positive side to this approach is that whether or not something is autonomous—or is autonomous as humans are—is a function of its internal organization. Whatever it is that enables humans to pass up a cupcake or work overtime: as long as machines have the functional equivalent then their actions will be autonomous too. There is, at least in principle, a way to decide if something is capable of autonomous action.

Despite the cleanness of this approach, it fails to incorporate many of the lessons that have been learned during the 4E revolution. Action, belief, emotions, desire, and intentions aren't obviously states that supervene on an internal substrate. Rather, intentional states and actions are processes that loop between organisms and their environments. Enactivists, for example, are quick to remind us that minds are *enacted* through interaction with the world. These insights about mind do not sit well with the approaches common to non-reductive physicalism.[12]

But another way to ask about what machine autonomy looks like adopts a social approach—a hybrid of sorts between Dennett's (1987) intentional stance and Strawson's (2014/1962) reactive attitudes. Dennett reminds us that human agents ascribe agency to a machine if it's the best way to make sense of that machine's behavior. For Dennett, we say of a machine that it "wants" or "thinks" if such intentional state ascriptions outperform other ascriptions in predicting the machine's behaviors. Strawson reminds us that others' intentions and attitudes towards us matter a great deal. Once we adopt long-standing habits of thinking and talking about machines in these ways, then Strawson's truisms about human relations begin to apply to AMAs: we care a great deal what others think of us,

[7] Thanks to Stephen Cowley for making this point explicit for me.

[8] See MacIntyre (1981) for critical discussion of 'ought implies can.' Note that his objection is that moral narratives can pull us in two directions; we ought to do two incompatible actions. MacIntyre doesn't consider the principle in light of the metaphysics of moral minds. My point here and his are consistent with one another.

[9] Kant, in "A Supposed Right to Lie Because of Philanthropic Concerns", says that we are not allowed to lie even to the murderer standing at our door, looking for a victim whose whereabouts we know. To lie to the murderer would be to fail to respect his autonomy and rationality. But see Langton 1992 for another interpretation of this case.

[10] 'Autonomy' can be predicated of people but also of mental states and actions. 'Autonomy' is said of people whenever their mental states and actions have the property of being performed autonomously: an entity is an autonomous agent when they believe, desire, intend, and act autonomously. To keep things readable, we'll talk about "autonomous action," but what is said here can *prima facie* apply to mental states as well.

[11] See Kim (1993) for excellent discussions.

[12] Despite what Clark and Chalmers (1998) would have you believe.

we are engaged in a variety of different relationships with different expectations, etc. The central point, Strawson (p. 7) tells us, of these truisms is,

> to try to keep before our minds something it is easy to forget when we are engaged in philosophy, especially in our cool, contemporary style, viz. what it is actually like to be involved in ordinary interpersonal relationships, ranging from the most intimate to the most casual.

An important question, then, is what our relationships with AMAs will be like: will they be to us as friends or servants or slaves? Will some of us fall in love with them? Will we care what they think of us? Will I bend over backwards to earn the respect of an AMA? Will I be jealous of an AMA's accomplishments? I doubt there are easy answers to these questions, but they highlight issues about AMAs fitting into the fabric of human life. If an AMA is autonomous in ways that are recognizable to us, then it cannot be merely an It but must in some cases be a Thou.

Intuitions about the significance of others' intentions and attitudes towards us is at the heart of Strawson's theory of reactive attitudes. An important theoretical fault line is between participant and objective attitudes. That is, when we treat people as full-blooded, intentional agents who are appropriate targets of responsibility-ascriptions and when we do not. We employ participant attitudes in the course of our everyday lives. When my neighbors continue to have loud parties after I've requested that they keep the volume down, I adopt an attitude of resentment. I am treating my neighbor as a cause of my annoyance and as disregarding my well-being. I am resentful because my neighbor, with full malice aforethought, ignores my request to keep their music down. I adopt a participant attitude.

But when my neighbor stumbles and then steps on my ingrown toenail, I'm not angry (though I'll be in a good deal of pain). The reason is because my neighbor bore me no ill-will in causing me pain; they didn't *mean* to do it. So while I'm upset and in pain, I don't adopt an attitude of anger or resentment at my neighbor. I adopt an *objective* attitude. I don't see them as disregarding my well-being. I see them as a cause for my pain but their causing my pain didn't involve thinking poorly (or not at all) of me. Strawson notes that we do this whenever someone does something by accident or when the person did not know what they were doing, either by biological constitution or circumstances. Currently, machines aren't deeply enough embedded in the right ways into human lives for us to adopt either stance. They are not targets of reactive attitudes of any type. I can be angry that a machine is not working as it should, but this is not to adopt a reactive attitude any more than when I'm angry that a thunderstorm has ruined my picnic.

The Dennett-Strawson approach sets the bar for ascription of autonomous action much higher than non-reductive physicalist approaches. How? It's not a matter of finding the relevant material substrate; rather, it's a matter of people accepting machines into the warp and woof of our lives. It's thinking about machines as Thou and not It. Given that some folks have trouble seeing others with different pigmentation as a Thou, I doubt non-organic entities will be widely enjoying these privileges anytime soon.

Autonomy is ascribed to AMAs when and only when we adopt reactive attitudes towards machines. When autonomy is ascribed to machines on this way of thinking, it will signify that machines are a part of the moral and social fabric of human lives. They will be afforded all the rights and privileges thereunto. We will truly be able to resent a drone for killing an innocent person; an AMA will experience, one would hope, guilt at having done it as well. We will be able to express gratitude for an AMA that sacrificed itself to save another. Creating an AMA would be no more or less morally problematic than having a baby.

But this approach comes at a cost. Prejudices might prevent machines from having intentions ascribed to them. Behaviors performed by an AMA, when performed by a human, might be subject to the ordinary range of reactive attitudes. But because the behavior was performed by a machine, it's not the subject of any reactive attitude. Prejudices towards AMAs will be like any other kind of race- or gender-based discrimination. AMAs would be incapable of acting autonomously because they would fail to be recognized as autonomous agents. But the similarity is not perfect. Those who would fail to ascribe psychological properties to a person on the basis of race or sex are demonstrably wrong. However, failing to ascribe psychological properties to a machine on the basis that it's a machine is not demonstrably wrong. The difference is that we know that dignity due to persons is in virtue of their being persons. Whatever it is that makes a human a person, it's not affected by race or sex. The same cannot be said right now for machines. It is not clear at this point that the thoughts and feelings of AMAs will be like ours.

The cost here is indeed a high one, but so are the stakes. AMAs are unlike anything we've seen in human history: creating machines that can think for themselves. If we're going to judge artificial agents by human standards, then those machines must be such that they could be incorporated into the fabric of human relations. The ball, then, is in our court to discern when AMAs are sophisticated and sensitive enough to become a part of our lives.

6 Genuine and perceived autonomy

So far, we've been entertaining AMAs having genuine, honest-to-goodness autonomy. But what we've neglected to consider, which is a theme of the special issue, is perceived autonomy. What exactly is perceived autonomy and how is it different from genuine autonomy?

One way to go about answering this question is to use an old philosophical distinction between perception and reality: a perceived autonomous moral agent (pAMA) is one that is believed to be autonomous by individuals but actually isn't. Much like the straight stick plunged partway in the water, how it appears is different from how it is. Right away, we can see (pun fully intended) that this approach doesn't work with the Dennett-Strawson approach we've identified in the previous section. Entities count as members of moral communities depending on whether we admit them to our communities, on whether they're recognized as genuine moral agents.

I propose instead that what characterizes pAMAs is whether our explanations of their workings involve Black Box Models (BBM). A BBM is a model that can make accurate predictions but whose workings are largely inscrutable to humans (Rudin 2019). Users are aware of variables and data going into the model and the algorithm used to weight the variables in the model, but how the algorithm generates those weights is opaque. So users know what's going in and what's coming out, but not on what's going on inside the machine. An important part of BBMs is that their models are not readily interpretable. What does it mean for a model to be interpretable? This is not the place to provide all conditions but it is easy to prime our intuitions. A linear model with a handful of variables is a paradigm of an interpretable model. It is clear what the variables are, and how changes to the variables affect the outcome. Uninterpretable models, by contrast, have many variables whose relevance and combination are, by definition, opaque. Deep neural networks are at least sometimes considered to be uninterpretable since they deal with many variables that are combined in a multitude of ways to make their predictions. Their opacity is often what makes them vulnerable to adversarial examples, i.e. precise perturbations to model inputs that cause the model to interpret the input as something other than it is—to classify a "stop" sign as a "yield" sign in autonomous cars, for example.

Suppose that a machine employs a BBM and also that the machine isn't well enough integrated into our moral community to count as a member and be credited as autonomous. Nonetheless, we might grant that the machine is *perceived* as autonomous but doesn't have autonomy conferred on it by the community. Why? Because it's a BBM, its workings are opaque to us and rather sophisticated: transparent models can be sophisticated or simple, but (all other things being equal) opaque models tend to be more sophisticated.[13] So our machine is generating outputs for complex inputs but how it's doing that is unclear to us. In some ways, the machine is working like humans do: we're largely predictable but just how we arrive at the decisions we make is the confluence of an enormous number of factors and considerations. But humans are at least sometimes scrutable: we can ask people why they did something and fit their response into a rational pattern of behavior. The same can't be done with BBMs. We might look at the underlying algorithms, but BBMs are, by definition, uninterpretable, so their activities and outputs cannot be fit into a rational pattern of behavior. Yet their outputs are at least sometimes surprising. So they are not simple machines but they aren't (and likely won't) be credited with full-blown rationality and autonomy. The best place for these devices is as pAMA.

"Perceived" autonomy on this account does not mean "seems genuine but is not." Rather, "perceived" autonomy is, "we're not sure how it's working." This resonates with how we're thinking about autonomy on the Dennett-Strawson account. Though people are complicated and often work from hidden motives, we mere mortals are at least able to fit actions into broader patterns of behavior. When we first read of Abraham preparing to sacrifice Isaac because God asked him to, we're shocked by the suggestion and his following through on it (at least until angelic intervention). But we make sense of his action in light of his fidelity to God, even if we ourselves could never do such a thing. By and large, we make sense of people's motives, even if we can't always puzzle them out. We can, in the language of MacIntyre (1981) and Bruner (1990) place their actions in the narrative of a tradition. But a pAMA is working from a dataset and a set of algorithms by which it develops a model that is largely inscrutable. It is autonomy is merely perceived because our best developers and engineers struggle to make sense of how it works. Genuine autonomy is ascribed to those individuals with constellations of intentional and moral states to which it's rational to adopt reactive attitudes. And we know (or at least know enough) about how people work for the ascriptions to be coherent. The same isn't true about pAMAs. They're pAMAs are so-called *because* of the "anonymity of assemblages that use kit of various kinds."[14]

[13] This is just a fact of the matter: why go to all the trouble to make a deep neural network for something that can be modeled by a simple linear equation?

[14] Many thanks to Stephen Cowley and Rasmus Gahrn-Andersen for this delightful and apt way of putting the matter.

This position resonates with Gahrn-Andersen's (this issue) thinking about perceived autonomy.[15] On his view, our engagements with technology always involve a pre-reflective experiential component. The categories of autonomy and heteronomy tacitly shape our engagements with things and people in the world. The category of heteronomy enables us to have an affinity for inert entities; it is in virtue of this pre-reflective category that we can engage in the existential structure of 'being-together-with'. And the category of autonomy enables us to have an affinity for other autonomous entities, which allows us to 'be-with' those others (ms. 6–7). Connecting these categories to the phenomenon of the 'uncanny valley', Gahrn-Andersen writes,

> Machines fall into the uncanny valley from a violation of affinity linked to a person's tacit understanding. Accordingly, one should be less interested in the objective traits of such machines than in how they actually appear to, and are perceived by, human subjects (ms. 9).

On my view, pAMAs' autonomy is merely perceived because their inner workings are inscrutable to even their creators. The fact that there are algorithms guiding their workings is no more relevant to an understanding of their functioning than the fact that there are neuroscientific principles undergirding our own activities. The objective traits, and the hype, of pAMAs do little to shape our perceptions of them by us. In the language of Gahrn-Andersen, pAMAs lack genuine autonomy because we cannot exist in the mode of being-with them. We cannot be-with them partly because we cannot put the activities of pAMAs into any sort of narrative.

7 Conclusion

I applaud White for raising a set of extremely important questions about the moral permissibility of creating KAMAs. I've suggested that the more pressing questions isn't whether they might be created morally but rather recognition of autonomous action in machines. I don't think that a functionalist approach will suffice; we have no idea if machine autonomy is going to work like human autonomy. Instead, I propose that we adopt conceptual resources from neo-Aristotelians like Dennett and Strawson coupled with insights from 4E cognition.

The American pragmatists urge us to ask what practical difference an idea makes. For supposedly autonomous machines, we have high stakes. To dream of creating KAMAs is to imagine machines as autonomous when they aren't. Or rather, it is to regard their decision-making procedures as subject to the same sorts of moral and epistemic norms as ours. If 4E theorists are right in thinking of cognition as embodied, embedded, enactive, and extended, then KAMAs' decision-making processes are not and cannot be like ours. Our bodies and lives are too different from what theirs would be. It's a difference in kind, not degree.

I've been largely critical of White throughout this paper, but I wholeheartedly agree with him on this: complex machines can be used to further *moral* ends and it is thereby an obligation to bring them into being. Machine learning, with sufficiently transparent processes, can make the world more just, not less. Treating these machines as heteronomous—until we subject them to reactive attitudes and they respond to them—acts as a bulwark against propping up our moral decision-making with processes that are alien to us.

We began by asking whether a robot could flirt. This test case primes intuitions about welcoming robots into our social communities as full-blooded autonomous agents. But things get serious quickly when we consider the manifold roles people play: police officer, judge, doctor, educator, caretaker. Our roles come with a bewildering variety of moral challenges. Determining the extent to which we want to permit machines into our social worlds as full-fledged agents is not something to do lightly. And before that, there are difficult conceptual issues to sort out on what it means for a machine to be autonomous in the first place. I propose to follow the methodological lead of Aristotle. We wait. We observe. We do our best. We hope.[16]

References

Bruner J (1990) Acts of meaning. Harvard University Press, Cambridge, MA

Chemero A (2009) Radical embodied cognitive science. MIT Press, Cambridge, MA

Clark A, Chalmers D (1998) The extended mind. Analysis 58(1):7–19

Dennett D (1987) The intentional stance. MIT Press, Cambridge, MA

Fodor JA (1987) Psychosemantics: The problem of meaning in the philosophy of mind. MIT Press, Cambridge, MA

Greenemeier L (2017) 20 years after deep blue: how AI has advanced since conquering chess. http://www.scientificamerican.com/article/20-years-after-deep-blue-how-ai-has-advanced-since-conquering-chess/. Accessed 8 Aug 2020

[15] Gahrn-Andersen and I are working in different traditions with different conceptual resources. Even so, I submit that, much as Merleau-Ponty observed of himself and Ryle, our work is not all that far apart.

[16] I'd like to express my gratitude to Stephen Cowley and Rasmus Gahrn-Andersen for their invitation to read and respond to White's paper "Autonomous Reboot." Their offer and subsequent feedback has been helpful beyond measure. I am additionally grateful to White for engaging in a dialogue about these issues. His insights about autonomous machines encouraged me to think more deeply about my own commitments. Finally, my eternal love and gratitude to Michele Lassiter for listening to me drone on about machine autonomy.

Irwin T (translator) (1999) Nicomachean ethics by Aristotle. Hackett Publishing Company, Indianapolis, IN
Kenny A (1979) Aristotle's theory of the will. Yale University Press, New Haven
Kim J (1993) Supervenience and mind: selected philosophical essays. Cambridge University Press, New York
Langton R (1992) Duty and desolation. Philosophy 67:481–505
Lassiter C (2016) Aristotle and distributed language: capacity, matter, structure, and languaging. Lang Sci 53:8–20
Lassiter C (2019) Language and simplexity: a powers view. Lang Sci 71:27–37
Lewis DK (1996) An argument for the identity theory. J Philos 63:17–25
MacIntryre A (1981) After virtue. Notre Dame University Press, South Bend
Putnam H (1975) The mental life of some machines. In: Mind, language, and reality: philosophical papers, 2. Cambridge University Press, New York, pp 408–428
Rudin C (2019) Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nat Mach Intell 1:206–215
Strawson PF (2014/1962). Freedom and resentment and other essays. Routledge, New York
Tonkens R (2009) A challenge for machine ethics. Minds Mach 19(3):421–438
Versenyi L (1974) Can robots be moral? Ethics 84:248–259
Vukov J, Lassiter C (2020) How to power encultured minds. Synthese 197(8):3507–3534