Shared Moral Foundations of Embodied Artificial Intelligence

Joe Cruz
Department of Philosophy and Program in Cognitive Science
Williams College
jcruz@williams.edu

Abstract

Sophisticated AI's will make decisions about how to respond to complex situations, and we may wonder whether those decisions will align with the moral values of human beings. I argue that pessimistic worries about this value alignment problem are overstated. In order to achieve intelligence in its full generality and adaptiveness, cognition in AI's will need to be embodied in the sense of the Embodied Cognition research program. That embodiment will yield AI's that share our moral foundations, namely coordination, sociality, and acknowledgement of shared resources. Consequently, we can expect a broad moral alignment between human beings and AI's. AI's will likely show no more variation in their values than we find amongst human beings.

Introduction

The ultimate ambition of artificial intelligence is to achieve in a made machine the capacity to cognize dynamically and adaptively in real world settings and in real time. The standard of cognition here is, of course, biologically evolved intelligence in us and in sophisticated nonhuman animals. We have as a goal the creation of beings that will thrive in open-ended environments, and we can readily create a list of the faculties that we want our artificial intelligence to have that includes rational thinking, memory, perception, creativity, forethought, abstract conceptualization, and perhaps even consciousness.

As we get closer to realizing the ambition of AI, we must come to grips with the moral (or immoral) potential actions of the intelligent beings we create. That is, we must consider what such beings will find morally desirable and morally permissible, especially concerning how to treat us and how to treat one another. We might wonder whether they will share a core set of human values, or— if we think human beings have no single universal morality—whether the morality of AI's will fall onto a recognizable landscape that will enable us to co-exist with them just as diverse human societies coexist.

It is my view that dystopian predictions and pessimism surrounding these issues of *value alignment* (Stuart Russell 2014) are overstated. There are, to be sure, countless possibilities for unintended consequences in the creation of beings that have their own wills. But I argue that the very project of anchoring AI's in our conception of intelligence—with all that that means in terms of embodiment, sociality, coordination and situatedness—implies that they will substantially share our moral foundations. AI's will be beings that we can make sense of and negotiate with just as we do when we find cultural difference in our own species. I am not claiming what is obviously false, that creating intelligence guarantees that those beings will be moral in the sense of doing mostly morally good acts. After all, human beings are not unfailingly moral. Nor am I denying that we can and probably will make smart special purpose devices that will serve immoral ends that we contrive. Those special purpose agents are not what I'm considering here.

Instead, my claim is any AI that seems to us genuinely intelligent in the full sense that human beings are intelligent will have to be embodied, and in so being will inevitably be governed by the same foundational considerations that generate morality it us.

Value Alignment Pessimism

It is reasonable to accept that someday AI's will be as intelligent as human beings are. Indeed, they may be more intelligent in the sense of facing fewer cognitive resource constraints—in memory, processing speed, and attention, for instance—and fewer sources of irrational distortion in their reasoning. Intuitively this leads to an open question concerning whether the superior reasoning of AI's will sometimes lead to decisions that a human being would view as morally unacceptable. The idea, remaining at the intuitive level, is that an optimizing reasoner may think

"coldly" and without regard for the things that are important to human being. The utility function that an AI has, itself the product of past decision making, may end up constituting an overall motivation that is alien and anathema to us.

This worry can be made more precise. For example, (Bostrom 2012) has argued that intelligence and final goals are orthogonal axes on which an AI can develop. An AI may achieve great intelligence without it being guaranteed that its final goals will be aligned with ours. Bostrom says,

> The orthogonality thesis implies that synthetic minds can have utterly non-anthropomorphic goals—goals as bizarre by our lights as sand-grain-counting or paperclip maximizing. This holds even (indeed *especially*) for artificial agents that are extremely intelligent or superintelligent (sect. 1.3).

His argument for the orthogonality thesis (OT) is that intelligence and motivation are not mutually entailing. The general form of Bostrom's argument appeals to mere possibility in that it is logically possible to pull apart particular ends from effective motivations. This, he acknowledges, recalls the Humean thesis that reason and desire are separate (Hume 1736), but Bostrom points out that OT can be correct even if Hume is not. In offering OT, Bostrom is rejecting metaethical views that propose a tight connection, perhaps even a constitutive connection, between motivation and apprehending the good through rationality. Since Bostrom characterizes intelligence in terms of instrumental rationality, and since it seems that in principle an artificial agent can lack whatever features of full blown rationality that make motivation and the good intrinsically intertwined, he concludes that there is nothing to prevent an AI—again, in principle—from being *mis*aligned in terms of its values.

Bostrom's argument is predicated on there being few constraints at the outset on how instrumental rationality or a utility function is instantiated in an AI. He later proposes *instrumental convergence*, namely the idea that, "…there are some *instrumental* goals likely to be pursued by almost any intelligent agent, because there are some objectives that are useful intermediaries to the achievement of almost any final goal" (sect. 2). Alas, this convergence does not tend to generate value alignment. Instead, it highlights some general trends in instrumental rationality that AI's and human beings will likely share, including, for instance, cognitive enhancement and resource acquisition.

The vulnerability of both the intuitive concern and arguments like Bostrom's is that they are hostage to factors that in fact constrain an intelligence such that its decision making will be aligned with other beings that face the same constraints. I speculate that the intuitive argument gets its force from the fact that relaxing cognitive resource constraints makes people think that *all* constraints—or, at any rate, all constraints that might encourage value alignment—are relaxed. This does not follow. Even if an AI is a faster reasoner with better memory access to relevant inputs to the argument and with fewer sources of irrational bias, there may yet be other kinds of constraints that funnel the decision making process into a commonality with human beings. I will offer just such constraints, below.

Likewise with Bostrom-type arguments. We can accept the in principle logical possibility that AI's need not be value aligned with us while we at the same time endorse that real world agents are very likely to be aligned. Thinking about the utility function as a rationality maximizer that is also a black box makes the mistake of leaving out the constraints on the function that are pre-rational. They are not *irrational* constraints in the ordinary sense of irrationality. Instead, they are the terrestrial, embodied conditions that are the background for any kind of rationality. Bostrom was getting at this with his instrumental convergence, but that convergence remains at the level of disembodiment. I therefore argue in what follows for a convergence that generates a more substantive alignment than Bostrom's mere instrumental convergence. My claim is that embodiment itself produces constraints on an intelligent being that will tend to push that being toward the same moral foundations that we have. In order to set up that claim, let us look again at the background image of OT and similarly abstracted characterizations of intelligence.

Isolated Cognition

By *isolated cognition*, I will mean intelligence that takes place within closed and idealized domains. It is, of course, perfectly understandable that early AI work would have ranged over seemingly tractable, well-structured environments like predicate logic, chess, or Tower of Hanoi puzzles. Worries about isolated cognition in this sense are very familiar and occurred even early on. The complaints are manifold: there is no assurance that the techniques used will scale up to real world settings, the successes achieved in such settings may be due to building in solutions specific to the domain so that there is no chance of achieving general intelligence, and the entire background inspiration—that intelligence can be achieved by adding together success at isolated problems, none of which are full blown intelligence by themselves—is contentious.

We should keep ready at hand the distinction between bounded rationality (Simon, 1957; Cherniak 1986) and isolated cognition. The critic of isolated cognition may still maintain that rationality is bounded. That is, there may be no sense in talking about rationality unless it is constrained by the real world limitations and constraints of actual cognizers. But being bounded by the realized resources at hand is a different issue than having intelligence be limited to

isolated, idealized domains. One is a point about the range of resources that the intelligent being can bring to bear and the other is about the range of challenges that an intelligent being is capable of meeting. I am concentrating here on the later, and my claim is that we all more or less accept that we have not created an AI if its cognition is isolated.

However, in spite of what I expect is widespread acceptance of this sort of critique, I maintain that there is a version of the isolated cognition problem that remains uncritically accepted. That version is where we tolerate a conception of AI that does not emphasize the *embodiment* of the intelligence. I mean explicitly to be referencing the research direction of Embodied Cognition as it is pursued in cognitive science, which I discuss in the next section. Even researchers who have internalized the critique of isolated cognition in AI still tend to accept disembodied intelligence as a tacit assumption.

Perhaps ironically, in popular culture the image of artificial intelligence is woven in with imagery of autonomous robots and artificial bodies. Maybe, then, the tight connection between being intelligent and having a body is more widely accepted in the folk image of AI than in the image of specialists and experts. But even in popular culture there is often the implication that being embodied is in a way optional for an intelligence. The intelligence needs to learn its alien body by practicing movement, or finds its body and the world of concrete objects just another intellectual object to think about and figure out. It is tempting to denigrate this as a Cartesian conception of intelligence, though Descartes himself thought that the human condition was a radical fusion between body and mind (Descartes 1641). Putting those historical niceties aside, just as it is sometimes believed that the mind can be understood as separate from the body, so too do AI researchers sometimes conduct themselves as if cognition can be realized without attending to the bodies of the cognizing beings.

Bostrom's OT is guilty of proffering isolated cognition. He says, "…'[I]ntelligence' will be roughly taken to correspond to the capacity for instrumental reasoning… . Intelligent search for instrumentally optimal plans and policies can be performed in the service of any goal" (sect. 1.2). This instrumental reason can take on any goal, as if the terrain of possible goals is as open ended as anything we can conceive. Notably, in this conception of intelligence no goals are delivered by the concrete realization of the body of the AI, or by the world of terrestrial or cultural objects that the AI inhabits. This, I urge, is intelligence of the most arid, ungrounded sort, and a tacit acceptance of the isolated cognition thesis. This conception of intelligence obscures the substantial and likely overlap of moral foundations between us and future AI's that will result from the fact that we must be and AI's must be embodied.[1]

We may still ask, what is embodiment? In the next section I briefly sketch some of the themes that fall under Embodied Cognition research. My goal is to make it plausible that, once an intelligence is embodied, it is constrained into the very same foundations of morality that we are. This suggests that the value alignment problem between us and AI's will be no greater than the value alignment problem between human cultures or individuals.

Embodied Cognition

Embodied cognitive approaches exhibit as wide a range and heterogeneity as more traditional computational approaches. Consequently, there is little that can be said by way of a general characterization of Embodied Cognition (EC) that won't have to be immediately qualified, elaborated, or downright retracted. Still, I think this much is right. EC seeks to place closer to the center of cognitive science at least the following five topics:

- The physical morphology of an intelligent being and its system of neurological control (Beer 1990; Thelen 1995; Chiel and Beer 1997).
- The possibilities of understanding cognition without appealing to internal explicit representations over which algorithmically driven computation takes place (Brooks 1991; Chemero 2009).
- The evolution of intelligence (Brooks 1991; Anderson 2003).
- The interaction and close coupling between the environment—sometimes including the cultural environment—and a cognitive agent (Beer 1990; Hutchins 1994; Clark 1997).
- The conscious phenomenology of having a body (Varela et al 1991; Thompson 2010; Nöe 2004; Toner et al. 2016).

Each of these conceptual threads of EC help push one in the direction of the following kind of thought: though it might have seemed that cognition, broadly construed, could be made sense of as disembodied and isolated from the environment, it turns out that there are sundry crucial ways in which thought is entangled with the condition of one's body. For example, (Hutchins 1995) argues that the cognition involved in ocean navigation critically involves the technical and cultural evolution of a much bigger system than an individual marine pilot. The instruments and

[1] A natural response here might be that it is possible, perhaps even likely, that AI's and future human beings will inhabit a virtual world within which there is no embodiment in the literal sense. Still, perceiving and cognizing the world *as if* embodied is equivalent to actual embodiment for the purposes of my argument.

body positions realized in their use, the traditions of information exchange between navigator and wider crew, and the internalized routinized actions of attention by an experienced navigator are all part of the computation taking place.

As another kind of example, consider that some kinds of central commands for actions are inert in the absence of naturalistic feedback from the environment. Leeches left to behave in a reduced preparation fail to produce functional swimming output because their swim interneurons do not fire at a high enough frequency. When placed in a full preparation with normal feedback from water, leeches then effectively swim (Chiel and Beer 1997; Kristan 2000). The water environment is required for the successful functional behavior of leeches. These and many other results persuade some cognitive scientists that intelligence cannot be made sense of in isolation from the body and environment (for an effective overview see Clark 2010).

The data and experiments that have driven EC are a scattered tableau from cognitive psychology, robotics, connectionist and dynamical approaches, cognitive linguistics, and the early 20th century history of phenomenology in philosophy. This variety can make it seem as if EC is struggling both conceptually and empirically. Indeed, a gloomy assessment of EC seems confirmed not least of all by a number of well publicized experimental results that served as common early encouragement for EC but that have failed to replicate. For instance, the facial feedback hypothesis involving a pencil gripped in the mouth were reported in (Strack et al. 1988), but have not been reproduced (Wagenmakers, E.-J., et al, 2016). It has been unhelpful, to put it mildly, for poorly designed psychological experiments to be treated as the load bearing support for EC. My view is that EC can be made sense of as a distinctive research direction in cognitive science separate from dominant trends in the field, and that there are good reasons both conceptual and empirical in support of EC. These research arcs should be viewed as that essential part of cognitive science that provides a corrective and reorientation of a maturing field.

It is tempting to try to assimilate the considerations of EC into an isolated cognition perspective. The idea would be to treat all of those bodily and environmental considerations as data or input to a central, isolated cognitive system that would then compute the right behavioral decisions. This, according to EC, is profoundly to miss the point. The body and the environment are in fact coupled with cognition in a fast and reliable system of feedback and calibration. To try to capture this coupling by representing the body and outside world such that those representations will need to be computed over is to incur a load and process that is far from the way natural intelligence is achieved. EC does not claim that doing so is metaphysically impossible, but that it is not the way evolved cognizers achieve their success. Representations, to the degree that they are had explicitly and on-board by the various individual agents, are no longer the primary focus of the analysis. Instead, we focus on the way that globally intelligible behavior emerges from interactions, and there may be no place in the system where it makes sense to say that a representation is explicit.

EC, then, can be seen as a methodological claim about cognitive science. It says that the proper unit of analysis for a model or theory of some particular cognitive behavior is a wide system that includes body and in some cases environment, and it makes the bet that modeling that wider system is more likely to lead to successes and insight regarding intelligence. This reading was always present in EC, but it tends to get obscured. The A-Life literature, for instance, attends to evolution and to the interaction between intuitively separate agents and is another historical source of inspiration for EC (Langton 1989; Wolfram 1994). We can see intelligence and cognition at the level of the agents that we intuitively recognize as individuals—birds or termites as the case may be—but we can also see intelligence at a different level, that of the collective that originates from the interactions between the individuals. That intelligence may be aimed at a different purpose and faces different constraints and therefore affords different opportunities to agents working together. Thus, the explanatory goals of cognitive science are shifted to embodied quantities.

Embodiment and Moral Foundations

The case that I am advancing asserts that pessimistic worries about the value alignment problem in AI originate in a tacit acceptance of cognition as isolated. Embodied cognitive research is a powerful corrective to viewing cognition as isolated, and shows that any being that we view as genuinely intelligent with respect to open ended problems is likely to be driven by its embodiment and coupling to the environment. But how does embodiment tend to make it such that an AI's values will be aligned with ours? In this section, I address this final step.

There is a literature on the role of body in moral cognition. The idea is that our moral judgments of, for instance, the rightness or wrongness of an action are grounded not in our rational, introspective reflections on the matter but instead in our bodily reactions at the time that we consider the action. For instance, it is alleged that our reaction of disgust is a primary driver in whether we think an action is wrong. (Haidt 1993) offers empirical evidence for the thesis that the way we morally judge the rightness of a family eating the family dog after it is killed by a car is based in our bodily and affective reactions. (Greene and Haidt 2002) likewise present empirical evidence that the appear-

ance of uncleanliness or the smells present when making a moral judgment substantively influence that judgment. These are cases reminiscent of the psychological literature on irrationality, where we are sometimes wrong about the causes of our beliefs (Nisbett, R. and Wilson 1977). In the moral domain, the particular incorrectness and the source of our judgment is something about our embodiment or our situatedness in a place.

It may well be that, when artificial intelligences are embodied, i.e., when they are bounded in their robotic casings and located in physical and cultural spaces alongside human beings and nonhuman animals, they will somehow rely on their bodily states to make moral judgments. This might align them with human beings who rely in part on affective states in our moral judging. On the other hand, it is an open question whether AI's as we envision them now can have bodily based emotions and feelings of, say, disgust. Moreover, it might be objected, couldn't we simply design AI's so that they didn't commit this sort of moral judgment irrationality and didn't let bodily states influence them? The reply to this possibility is that, if AI's don't employ bodily reactions in moral judgments, we may not count them as intelligent in the broad sense of being calibrated to a world that includes our own moral reactions. They will seem "off" in the moral domain and will not, more or less deeply, make sense to us. So I think it is likely that any AI worthy of being called intelligent with either have to have such bodily reactions (or simulations of them) and will have to use them in morally judging.

This is one way in which embodiment is relevant to an AI's moral behavior, namely the behavior of judging actions as moral or not. But this is a fairly modest allegation; I wish to go much further. I want to say that we can expect the moral horizons of AI's to be in substantial alignment with ours.

In order to determine whether an AI's values will align with ours in a more complete sense, we must consider what makes our values the ones that they are. The debate in philosophy on the source of values is as immense as any literature in the field, so it would be absurd to try to engage it here in any serious way. I am hoping, then, not to have to. In talking about the foundations of morality—and about whether those foundations will produce values that are in alignment between AI's and human beings—I am not referring to the level of metaethical theorizing that pits accounts like utilitarianism, virtue theory, social contract theory, or Kantian deontology against each other. I am gesturing at a deeper foundation that make it possible for those metaethical theories to be in the running, so to speak. The conditions will then be the foundational bedrock of whatever values develop. I propose at least these:

- Resources are finite. In an imaginary world where there are no limits on resources, virtually any desire can be met and so any utility function can be satisfied. That is a world where values are not required because no ordering of preferences is necessary. So, in order to get the question of values off the ground, resources must be finite.
- Coordination and collaboration with other intelligences is required to achieve the maximum flourishing of any individual. In a world where an individual intelligence can by itself achieve all of its desires, there is no need for values in the relevant sense, because the individuals will not need to be value aligned.
- The world is substantially composed of the scaffolding of tools, built environments, collective knowledge, and culture. These create the imperatives to achieve alignment. In a world where individuals form their own systems of tool use, knowledge communities, and cultures, there is no pressure to achieve values that must be aligned. This is because each individuals scaffolding can express that individual's particular values.

All three of these are impossibly abstract, but I hope that they provide the hints that I intend. These are descriptions of the conditions that make it so that human morality develops. Acknowledgement of shared resources, between-being coordination, and sociality are the foundations of morality in the sense that the need and nature of morality emanates from them. Moreover, all of those foundational dimensions have to do with the fact that we are embodied beings in an environment. We are not isolated cognizers who deploy a utility calculus that is pristinely detached from the world. Any AI that meets our intuitive and theoretical standards for intelligence will need to be embodied, and embodiment will establish AI's into our values because they originate in those moral foundations which in turn stem from embodiment.

But how, precisely, can these deeply abstract foundations align AI's behavior with our morality? Here is a sketch of the account. Finiteness of resources generates a shared morality in that all agents are working against the same backdrop and therefore must manage goals that potentially conflict in terms of resource use. If an agent has a goal that requires some amount of a resource and another agent also requires that resource, then there must be a resolution. That can be literal conflict or deception or division of the resource or a change in goal, to name just a few possibilities. What the possibilities have in common is a reckoning of other agents. Other agents must be treated in some way or other. Another way of putting this is that a precondition for instrumental reasoning with finite resources is the capacity to reckon with the instrumental reasoning of other beings. That creates the need to arrive at answers to value question, especially ones in the domain of reciprocity, fairness, and retribution.

With respect to collaboration and cooperation, I claim that this is a foundation for a shared morality in that it requires that agents acknowledge one another as beings *with*

whom to pursue goals, and as noteworthy beyond merely being objects or obstacles in the environment. Our intelligence is deeply bound up with our sociality, from the mundane observation that we solve many problems through a group cognitive effort to the more exotic point that our cognitive potential is massively extended by epistemic actions. These are actions that place in the environment information that is required to achieve a goal (Kirsh and Maglio 1994; Anderson 2003), and this is most effective when done collectively. For example, writing is probably the most crucial epistemic action that human beings undertake for our modern thriving. Recovering information from writing requires set conventions and systems of acting together that once again force agents to exist within a world where there must be coordination.

Go back to the idea of intelligence in general. If we focus just on effectively executing a utility function, we do not capture what we think of as distinctive about our own intelligence. The case of dogs is illustrative. Bonobos, dolphins, elephants, and African Grey Parrots each show serious cognitive advantages over *canis familiaris*. On a certain abstract, isolated conception of what intelligence is, it seems that putting dogs near the top is something only a starry-eyed dog lover could do. However, dogs are in another sense especially intelligent, and this makes perfect sense because they inhabit our coordinated, built, social world. To be sure, they inhabit our world due to a combination of natural and massive artificial selective pressure. They were probably at the right place at the right time alongside early humans. Whatever the origin, my view suggests that dogs are second only to us in being intelligent in an open ended, authentic way. Dogs possess something crucial to intelligence that includes a responsiveness to emotions, comfort and facility in built up human spaces, deep understanding of coordinative sociality, and eye contact referencing and regard for human beings as principal objects in the universe (Hare and Tomasello, 2005; Topal et al, 2014). So, too, must any AI worth the name. Genuine AI's will presumably add a substantial capacity for reasoning, abstraction, and language that dogs are not capable of, but that should not obscure the shared background. If this is correct, then a successful AI must inhabit our social world, which will require entering into a central dimension of it, our moral landscape.

Finally, the scaffolding of culture and the technologies we have produced have moral considerations as preconditions for membership. Being in a culture includes understanding how to act in that culture. Those behaviors, especially the ones that are embodied rather than explicitly theorized about, are what make an agent a natural, companionable co-intelligence in an historical and cultural milieu. Of course I am not claiming that AI's by their embodiment will automatically have the morality of my culture or any other specific culture. Indeed, we should be worried that engineers and programmers will make AI's too much members of their own culture and will thereby replay a kind of colonialism. But in order to be viewed as intelligent and by dint of being present with a body in a culture, an AI will need to be ready and receptive to cultural norms of co-existence that we identify as morality, whatever they may be in a given culture.

So, when Bostrom claims that an AI can pursue arbitrary goals (within the horizon of instrumental convergence) and that therefore there can be a problem of value alignment, my reply is that intelligence that is too far beyond the pale will not seem like intelligence to us. While such an AI might seem to have a goal directed cleverness the way that some non-human animals are clever, it will not seem like the kind of open-ended intelligence that we associate with, say, dogs. An AI's goals will be constrained by the fact that they share a world with us, and will enter into our technologies including and especially culture.

The view I am defending here has an affinity with moral foundations theory (MFT) proposed by social and anthropological psychologists (Haidt 2012). MFT alleges six foundations for human morality—care, fairness, loyalty, authority, sanctity, and liberty—and claims that these are encoded in our quick-to-deploy intuitions about moral cases. For instance, suffering as a basis for moral imperatives is alleged to derive from the evolution of parental care, and is characterized (Haidt and Craig 2004) as having compassion as one of its characteristic emotions (p59). This original trigger then culturally evolved to treat kindness as a virtue. MFT theorists attempt to bolster the considerably speculative cast of their claims with cross-cultural empirical data on moral intuitions and work from evolutionary psychology.

MFT is a descriptive theory of our moral judgments. It is not normatively recommending that this is the way that we ought to be toward one another. Whether we view these as a rationally sound normative ground is up for considerable debate. If we do not but MFT is descriptively correct, we may find ourselves in the regrettable position of rationally viewing patterns of conduct as immoral that we routinely are unable to correct because we are constrained by MFT type foundations. I think this would invite an interesting question of whether future AI's will be vexed by *our* lack of moral alignment with *them*. After all, there is no engineering necessity that dictates that AI's will be initially helpless and require literal parenting. Therefore, MFT's specific foundations may not be the ones that ensure that AI's fall within human moral horizons. AI's may be able to achieve something closer to a rationally ideal moral behavior, and may find us more subject to the biases countenanced by MFT.

That possibility, though, postulates that AI's are not constrained by something more fundamental than the

foundations proposed by MFT. Mine claim to be more basic.

Conclusion

To say it again, I am not claiming here that we are somehow obligated, morally or otherwise, to build beings whose moral views are substantially like ours. That would be a point about us and our design goals, and that is a different argument. My thesis is that, in setting out to build an intelligent being, we will automatically—as a kind of conceptual inevitability—create something that shares with us a fundamentally moral outlook because those beings will be embodied as we are and will inhabit our world as we do.

The moral outlook that AI's share with us may not determine specific behavior. Our AI's CPU won't explode if it tries to do something immoral. But then again sharing a moral foundation does not obligate human beings in specific ways, either. Individual human beings routinely do immoral things. Further, AI's won't, on my argument, be ultramoral, though we could probably build them to be that way. I suspect that such ultramoral AI's will also seem "off" to us. They would likely fail to exhibit a certain pragmatism or contextually appropriate selfish behavior and will therefore seem not quite intelligent.

AI's will be aligned with us in the same way that human beings are aligned with one another. Human beings are impelled to be moral by a complex of conscious and non-conscious forces that are a response to their embodiment. Above, I have tried to gesture at a way of conceiving of those forces. They constitute more of a constraint on value alignment than has been appreciated.

References

Anderson, M. 2003. Embodied Cognition: A Field Guide. *Artificial Intelligence*, 149 (1):91–130.

Beer, R. 1990. *Intelligence as Adaptive Behavior: An Experiment in Computational Neuroethology*. San Diego: Academic Press.

Bostrom, N. 2012. The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents. *Minds and Machines* 22(2), May:71-85.

Brooks, R. 1991. Intelligence without representation. *Artificial Intelligence*. 47(3):139-159.

Chemero, A. 2009. *Radical Embodied Cognitive Science*. Cambridge, MA: MIT Press.

Cherniak, C. 1986. *Minimal Rationality*. Cambridge, MIT Press.

Chiel, H., and Beer, R. 1997. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in Neuroscience* 20(12):553-557.

Clark, A. 1997. *Being There: Putting Brain Body and World Together Again*. Cambridge, MA: MIT Press.

Clark, A. 2010. *Supersizing the Mind: Embodiment, Action, and Cognitive Extension*. Oxford UP.

Descartes, R. 1641/2015. *Meditations on First Philosophy with Selections from Objections and Replies*. Cottingham, J., ed. Cambridge: Cambridge UP.

Gallese, V., and Lakoff, G. 2005. The brain's concepts: The role of the sensorimotor system in conceptual knowledge. *Cognitive Neuropsychology* 21:455–479.

Greene, J., and Haidt, J. 2002. How (and where) does moral judgment work? *Trends in Cognitive Sciences* 6(12):517–523.

Haidt, J.; Koller, S.H.; and Dias, M.G. 1993. Affect, Culture, and Morality, or Is It Wrong to Eat Your Dog? *Journal of Personality and Social Psychology* 65(4):613–628.

Haidt, J.; Craig, J. 2004. Intuitive ethics: how innately prepared intuitions generate culturally variable virtues. *Daedalus* 133(4): 55-66.

Haidt, J. 2012. *The Righteous Mind: Why Good People are Divided by Politics and Religion.* New York: Pantheon Books.

Hare, B., and Tomasello, M. 2005. Human-like social skills in dogs? *Trends in Cognitive Science* 9:439-444.

Hume, D. 1736/1978. *A Treatise of Human Nature*. Selby-Bigge, L.A., ed. Oxford: Oxford UP.

Hutchins, E. 1995. *Cognition in the Wild*. Cambridge, MA: MIT Press.

Kirsh, D., and Maglio, P. 1994. On distinguishing epistemic from pragmatic action. *Cognitive Science* 18(4): 513-549.

Kristan, W. et al. 2000. Biomechanics of Hydroskeletons: Studies of Crawling in the Medicinal Leech. In *Biomechanics and Neural Control of Movement*, Winters, J. and Crago, P., eds. Springer-Verlag.

Lakoff, G., and Johnson, M. 1999. *Philosophy in the Flesh: The Embodied Mind And Its Challenge To Western Thought*. New York, NY: Basic Books.

Langton, C., ed. 1989. *Artificial Life*. Redwood City, CA: Addison-Wesley.

Noë, A. 2004. *Action in Perception*. Cambridge, MA: MIT Press.

Russell, S. 2014. Of Myths and Moonshine [Blog post]. Retrieved from https://www.edge.org/conversation/the-myth-of-ai#26015.

Simon, H. A. 1957. *Models of Man, Social and Rational: Mathematical Essays on Rational Human Behavior in a Social Setting*. New York: John Wiley and Sons.

Strack, F.; Martin, LL.; and Stepper, S. 1988. Inhibiting and facilitating conditions of the human smile: a nonobtrusive test of the facial feedback hypothesis. *Journal of Personality and Social Psychology* 54(5): 768-77.

Thelen, E. 1995. Time-scale dynamics in the development of an embodied cognition. In *Mind In Motion*, Port, R. and van Gelder, T., eds. Cambridge, MA: MIT Press.

Thompson, E. 2010 *Mind in Life*. Cambridge, MA: Belknap Press of Harvard UP.

Toner, J.; Montero, B.; and Moran, A. 2016 Reflective and Prereflective Bodily Awareness in Skilled Action. *Psychology of Consciousness: Theory, Research, and Practice* 1:1-13.

Topál, J.; Kis, A.; Oláh, K. 2014. Dogs' sensitivity to human social cues: A unique adaptation. In *The Social Dog: Behavior and Cognition*, Kaminski, J. and Marshall-Pescini, S., eds. Sand Diego: Elsevier, pp. 319-346.

Wagenmakers, E.-J.; Beek, T.; Dijkhoff, L., et al. 2016. Registered Replication Report, Strack, Martin, Stepper (1988). *Perspectives on Psychological Science* 11: 917–928.

Wolfram, S. 1994. *Cellular Automata and Complexity*. Redwood City, CA: Addison-Wesley.

Varela, F.; Thompson, E.; and Rosch, E. 1991. *The Embodied Mind*. Cambridge, MA: MIT Press.