Studies in History and Philosophy of Biological and Biomedical Sciences 44 (2013) 679–687

Contents lists available at SciVerse ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

journal homepage: www.elsevier.com/locate/shpsc

Machine wanting

Daniel W. McShea

Dept. of Biology, Duke University, Box 90338, Durham, NC 27708-0338, USA

ARTICLE INFO

Article history:
Available online 21 June 2013

Keywords:
Emotion
Teleology
Purpose
Goal-directedness
Artificial intelligence
Robot

ABSTRACT

Wants, preferences, and cares are physical things or events, not ideas or propositions, and therefore no chain of pure logic can conclude with a want, preference, or care. It follows that no pure-logic machine will ever want, prefer, or care. And its behavior will never be driven in the way that deliberate human behavior is driven, in other words, it will not be motivated or goal directed. Therefore, if we want to simulate human-style interactions with the world, we will need to first understand the physical structure of goal-directed systems. I argue that all such systems share a common nested structure, consisting of a smaller entity that moves within and is driven by a larger field that contains it. In such systems, the smaller contained entity is directed by the field, but also moves to some degree independently of it, allowing the entity to deviate and return, to show the plasticity and persistence that is characteristic of goal direction. If all this is right, then human want-driven behavior probably involves a behavior-generating mechanism that is contained within a neural field of some kind. In principle, for goal directedness generally, the containment can be virtual, raising the possibility that want-driven behavior could be simulated in standard computational systems. But there are also reasons to believe that goal-direction works better when containment is also physical, suggesting that a new kind of hardware may be necessary.

"What would you like to do this afternoon?" Not a machine in the world can honestly answer that question. No computer yet built can give an answer and mean it. It would be easy to give a machine a list of possible answers and a random number generator: Sit on the couch sipping tea and eating bonbons. Surf the web. Plow the "back 40." Then flip a (three-sided) coin. Or we could give the machine some sensitivity to the world, letting it make a choice based on small differences in external variables, such as the present condition of the house, the time since the last web surf, and the weather, with positive and negative weightings assigned to each input variable, themselves perhaps based on positive and negative results of past decisions. But by whatever algorithm it decides, it won't really *want* to sit on the couch, surf the web, or plow the land behind its farmhouse, the back 40. It will have no real *preference*. It won't *care* whether or not it is able to do what it chooses.

In our struggle to understand human thinking and to replicate it in machines, wanting-preferring-caring ought to be central, more central than reasoning. Wanting-preferring-caring should also be more central than the emotions (to which they are interestingly related but still different from). Wanting-preferring-caring is the cause, and the only possible cause, of all deliberate thought, speech, and action. It is the seat of agency in humans, and in every other species that behaves deliberately. It is the motive force that drives deliberate thought, speech, and action (in other words, "behavior"). No serious simulation of human-style interaction with the world is possible without it. My positive point here will be that simulating wanting—machine wanting—is possible, or at least, there are no known barriers. However, it has not been done. Perhaps in part because no one has tried? (To my knowledge, it has not been tried, but this is not my field.) Anyway, the main problem seems to me to be that little is known about how wanting works in animal/human minds and brains. More precisely, I should say, much is known about the various neurons and brain regions *involved* in this or that kind of choice, about the various psychological factors that *affect* this or that sort of preference (Dolan & Sharot, 2012), but little is known about the physical structure and dynamics that corresponds to wanting-preferring-caring, about what wanting-preferring-caring *is*. Thus, we do not even really know what we would be trying to simulate. In this near

E-mail address: dmcshea@duke.edu

1369-8486/$ - see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.shpsc.2013.05.015

understanding-vacuum, it would be quite bold to propose a full recipe for how to proceed. I am not that bold. Instead, I offer two basic principles that I believe should guide future thinking about the problem. The first is that all wanting is non-rational, non-logical, and since our best "thinking" machines now are logic machines, we must—if we hope to get them to want—build them along fundamentally different lines. The second is that wanting is necessarily teleological, and this fact tells us something useful about the structure of any system that wants.

To explain these claims, I need to do the impossible. I need to overthrow a habit of mind that has become standard in some academic circles. That is what I need to do just to *explain* these claims. To *convince* you of them is doubly impossible. But two things give some meager cause for optimism. First, the view I propose is intuitive, mirroring as it does the standard folk psychological story about how wanting works. And second, it follows an old and distinguished line of argument in philosophy, beginning with Hume. I begin with Hume.

1. Reason and passion, logic and wanting

"Nothing is more usual in philosophy, and even in common life, than to talk of the combat of passion and reason . . .," writes Hume in the second book of *A Treatise of Human Nature* (Hume, 1740 [1978], p. 413). He goes on to explain why such talk is nonsense. If Hume is right—and he is—how can we still talk this way, centuries later? The psychologist Jonathan Haidt, in his recent book *The Happiness Hypothesis*, compares passion to an unruly elephant and reason to its rider. Our lives, says Haidt, are a constant struggle by the rider to control and guide the elephant. Every modern reader instantly understands the point of the metaphor: reason versus passion.

Meanings shift with time and context. Hume and Haidt are actually talking about very different things. By "reason" Hume meant what we mean by "logic." By "passion" he meant roughly what I earlier called wanting-preferring-caring. And his argument in the *Treatise* is that logic, and the conclusions of logical reasoning, have by themselves no motive force, and therefore no power to control or guide or even nudge our wants. Let me say this again, this time more nakedly. His argument is *not* that logic has very little power to influence our wants. It is *not* that the force of logic is weak in comparison to the power of the wants. It is that logic has exactly zero power of influence. And the reason is that logic and wanting occupy different and incommensurate categories of mental phenomena. Logic is concerned with relationships among ideas and their representations, for example the relationships among numbers in mathematics and the relationships among ideas about objects in physics. Given Newton's second law, and an object with mass, logic tells us what we can say about its acceleration when it is acted upon by a certain force. Or, in the world of everyday ideas and representations, if the kid in the third row is a normal fifth grader, and if what I think I know about fifth-grade psychology is correct, then I can claim with all the authority of logic that his furtive looks and hand movements under the desk are attempts to secretly attach and store there the gum I saw him chewing in violation of school rules moments before. That is logic.

In contrast, wanting is a species of volition. A want is an urge, an impulse, a motivation. It may be an urge to act, but it could also be an urge to speak or to think. Preferring is closely related. A preference for this rather than that is a kind of urge, an inclination toward these ideas, words, or acts, rather than those. Caring is different but similar. It is perhaps a less specific form of wanting and preferring. I do not know exactly what I want to think, say, or do in this situation, nor even which sort of result I prefer, but I know that I care, that what I do will matter to me, perhaps in ways I cannot articulate (yet?), even to myself.

It is with some diffidence that I offer my understanding of these terms for affective states. Their meanings are poorly constrained, even in the psychological literature. And the street usages are even less constrained, to the point that is doubtful that any randomly chosen pair of people will understand them in even roughly the same way. I actually think there are good reasons for the vagueness of these words. The reasoning mind is every moment of the day immersed in a sea of affect, of wants, preferences, and cares. And like any small thing immersed in a big thing, like a worm immersed in an ecosystem, its view of the big thing is always partial. If it is able to grasp the whole at all, it does so only vaguely. In any case, my hope is that through repeated usage of these affective terms in a variety of contexts, the reader will get the gist of what I mean. And for present purposes, the gist is sufficient.

Hume's point is that no want, no preference, no caring, lies at the end of any chain of pure logic. Physics and physiology may tell me that the bus bearing down on me will, if it hits me, smash me to pieces, but it does not follow as a matter of logic that I should want to get out of the way, or that I should care whether or not I do. In fact, I do care. And that caring takes the form of the fear that I feel in the moment as I see the bus bearing down on me. It is also the more considered desire to be alive that I might feel on reflection after dodging the bus. But this caring follows the sight of the bus because of how my brain is structured, not as a matter of logic. Thunder follows lightning as a matter of physics, not as a matter of logic.

Likewise, my experience and understanding may tell me that my fifth-grade student is trying to hide his chewing gum, but no wanting-preferring-caring follows from this as a matter of logic. It does not follow that I, the teacher, want to scold him or punish him or even ignore it. It does not follow that I care what he does with the gum. I observe his movements, consult my experience, and apply logic to infer what he is up to. If I care about the result of that logic—about the conclusion that he is trying to hide it and chew it later, in violation of school rules—it is for reasons having to do with my affective organization, my motivational structure, not logic.

A misunderstanding is possible here owing to the dual meaning of the word "follow." Observing the student trying to hide the gum may *evoke* in me a motivation to act, and this evoked motivation might be said to "follow" from the observing. But the "following" here is following in time, and in a physical sense. My motivation to act follows—is physically caused by and therefore follows in time—my seeing, and thinking about what I have seen. And it does so owing to some unknown physical pathway in my brain. But following in this sense is very different from following in the sense of logical entailment.

So no want, preference, or care lies at the end of any chain of logic. Further, Hume argues, no chain of logic can oppose a want, preference, or care. Again the reason is that logic is a relationship of ideas, or in modern terms, of propositions. And thus it has no motive force. In modern terms, we might say that a want is a physical thing, or a physical process, one that exerts a force. And that is why only another want can oppose a want. Only a force can oppose a force. (One might argue here that wants are qualia, or are closely associated with qualia, and therefore not able to generate any force, but the assumption here, and in Hume, is that they are not *only* qualia, that they are also efficacious.)

Nor can a want, preference, or care contradict logic. Contradiction, Hume writes, is a disagreement of ideas or of representations of ideas. And a want is not an idea or a representation. It is, in his words, an "original existence," one that "contains not any representative quality." In modern terms, we might say that a want is a state of mind or of a brain. It is a thing. And a thing is not an idea or a representation. And therefore there can be no disagreement between a want and an idea, any more than there can be a