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

disagreement between a yardstick and the idea of 36 inches. My yardstick might be off. It might not be 36 inches long. And in that case the *proposition or claim* that it is 36 inches contradicts the claim that it is a true yardstick. But that is a contradiction because a measured length is a number, an idea, and a claim that the object is a yardstick is also an idea. Contradiction is a relationship between ideas, so there is no problem. In contrast, the yardstick itself is a thing. And things are "original existences" and so cannot contradict anything. Likewise a want is a thing and therefore cannot contradict anything. Although it can of course physically oppose—apply a force against—another want.

Or maybe we should think of a want as a physical occurrence, an event in the mind or brain, rather than a thing. But events are not ideas or representations either. And thus they cannot contradict claims. A claim about an event can contradict another claim about it. My claim that my team has won is contradicted by the posted official score, by the claim of the officials. But my celebration of the claimed victory is just an event, and contradicts nothing. It may be inappropriate, perhaps revealing my disconnect with reality, but there is no logical contradiction in my celebration.

Puzzles arise for the modern reader. First, Hume might seem to be claiming that wants and preferences—including something as basic as a preference for survival—are irrational. But in fact he is not. Rather he is claiming they are non-rational or extra-rational, not the sort of thing that could be rational or irrational. They are states or events, like rain, that occur for good physical reasons but not for rational or irrational ones. Is the formation and falling of a raindrop rational or irrational?

Second, Hume's argument that there is no logic to my preference for survival might sound wrong because survival is necessary to my having preferences in the first place. That is true. Presumably my preference for survival arose from a history of natural selection acting on my ancestors (tweaked on shorter-timescales by my social environment). Those with a stronger preference for survival had more offspring than those with a weaker one, or something like that, with the result that in certain fearful situations their descendants today experience a strong preference for survival. Let us say this story is true. Still it does not contradict Hume. The evolutionary story gives the causal history of the preference I experience, and that causal history explains in biological and physical terms why I *do* in fact prefer survival. But it doesn't explain why I *must*, as a matter of logic. I can imagine a world in which no individual has a preference for survival, say a world in which aliens create a series of generations of individuals with ever stronger propensities to commit suicide, a bizarre and thankfully improbable world indeed, but no violation of pure logic.

1.1. Modern reason and passion

Does it follow from Hume's argument that Haidt's metaphor is nonsense? If Hume has shown that reason can never oppose passion, then the rider cannot even influence the elephant, let alone control it. The elephant metaphor would seem to be nonsense, because reason opposing passion is a category mistake.

But that is true only under Hume's interpretation of reason and passion, reason in the sense of pure logic and passion in a sense that encompasses wanting-preferring-caring. And today we typically use reason to mean something more like "reasonableness" than like pure logic. We use reasonable to describe a normally motivated person at a moment when she is calm and thoughtful, when she is thinking through her situation, examining consequences of alternative actions, and bringing to that examination the full range of her wants-preferences-cares, weighted according to their importance in a long-run view of her life. We say that the thought and action of such a person are, at that moment, driven by reason. In contrast, we use passion only for *strong* wants, preferences, and cares. We describe as passionate a person at a moment when she is somewhat out of control, perhaps temporarily under the spell of one or a small number of strong or even overpowering wants, preferences, and cares. Under these interpretations, Haidt's story makes sense. And there is no category mistake in claiming that reason in this sense—infused as it is in this sense with wanting—can oppose passion, understood as very strong wanting. Reason in the modern sense has an element of what Hume called "calm passion." And calm passion can oppose strong passion. (In coalition with other calm passions, it can even win.) With reason and passion understood this way, the rider can influence the elephant.

1.2. Emotion

I pause here to make one small point about the modern study of emotion, which Hume's argument puts in a new light. Hume's notion of passion encompasses all wants, preferences, and cares, both weak and strong. For him, wanting an ice cream cone and fearing a snake are both "passions." Thus, ordinary wanting-preferring-caring are passions, just weak ones. And anger, fear, disgust, and strong sympathy are also passions, strong ones. But our modern ontology makes a different separation, between on the one hand the strong affective states, what we call the emotions, and on the other hand all other conscious mental states and processes, including both weak affective states (wanting-preferring-caring) and logical reasoning. In other words, the modern view implicitly lumps logical reasoning together with wanting-preferring-caring, treating both as non-emotional. From one perspective, this makes sense. There has been a fruitful line of research over the past century or so in biology, psychology, neurobiology, and lately economics, a line of research that is very interested in emotions as eruptive states. It has been interested in the evolutionary history of these eruptive states, the physiological arousal they accompany, their expression, their effects on memory and cognition, the biases they introduce in reasoning, their moral dimension, and so on (e.g., Ariely, 2010; Darwin, 1872; Haidt, 2006; James, 1890; Simon, 1967). In this context, it seems natural to study the strong affective states and processes to the exclusion of the weak ones, to study anger, fear, disgust, sympathy, etc., while either ignoring wanting or treating it as a wholly distinct phenomenon. But as Hume's argument reveals, there is also something odd and conceptually confused about this move. Studying emotions without studying wants, preferences, and cares is like studying the physics of supernovae while ignoring the physics of ordinary stars.

The point is this section is mainly negative. No chain of logic can conclude with a want, preference, or care. And therefore no pure-logic machine will ever want, prefer, or care. If we want our machines to do these things, if we want to simulate these things in a machine, we will need to structure our machines in some other way.

2. The structure of wanting

The incommensurateness of logic and physical things or processes is indisputable. In contrast, the claims of this section and the section that follows are highly disputable. They rest on a new (and still evolving) view of teleology, a view based on the theory of compositional hierarchies (Simon, 1962; Salthe, 1985, 2009; Wimsatt, 1974, 1976, 1994). This view was advanced in a recent paper (McShea, 2012) and is sketched here.

Wanting-preferring-caring arises physically, that is, in systems with a particular physical structure. We do not know much about that physical structure in people or other animals, or about what its structure would have to be in a machine version of

wanting-preferring-caring. But we do know something. We know that wanting has a teleological quality. It is goal directed, valenced, seeking, purposeful. My proposal here is that all teleological systems have a common structure, a hierarchical or nested structure. Nesting is often physical, consisting of a small thing nested inside a big thing (McShea, 2012). It can be virtual, consisting of an abstract point nested within a phase space. But when it is virtual, the physical structure that produces the phase space is typically physically nested anyway. If this line of thought is right, then wanting-preferring-caring almost has to involve a physically nested structure.

2.1. Persistence and plasticity

Consider a system that is *not* teleological. A Lego "Mindstorms" robot is a machine used to teach computer programming to middle and high school students. It moves on wheels, powered by a small electric motor. When activated, the robot executes a series of instructions programmed by the student into its on-board hardware, directing the motor and steering mechanism to move it forward, backward, turn left, etc. In one programming exercise, the student devises a series of instructions that direct the robot along a series of linear pathway segments, from one chosen point—marked with an X of tape on the floor—to a target X somewhere else in the room. So, for example, from the starting X, the robot might be programmed to move 1 meter forward, turn left, move 50 centimeter forward, turn right, etc., eventually arriving at the other X on the floor, the target. If the programming is done right, the robot will stop right on the target X, and in principle it should do so in every run.

Watching a few runs of this robot, its behavior might seem to have a goal-directed flavor. It is not hard to think of it as seeking the target X. But after watching many runs, it becomes clear that something is missing, something that emerges clearly whenever the robot goes awry. If, say, a student accidently nudges the robot while it is underway, or if it runs over a rough spot on the floor, causing it to leave the planned trajectory, it does not return to that trajectory. There is no error correction, and robot simply ends up somewhere other than the target. In the terms I will use here—borrowed from Nagel's (1979) treatment of teleology—the system has no persistence. It does not respond to perturbations by returning to its original trajectory. The robot will also miss the target if one changes the starting point. In Nagel's terms, its behavior is not plastic. It does not find the target from multiple starting points. These properties—persistence and plasticity—are the signature of teleological systems, Nagel argues. A moth seeking light shows persistence and plasticity. So does a sound-tracking torpedo closing in on a target ship. So do the cells of a developing organism, as it transforms from embryo to adult. Indeed, so would a different sort of Lego robot, one that uses a GPS to navigate to the target.

2.2. Upper direction and lateral direction

How do teleological systems do it? How do they achieve persistence and plasticity? The answer is that they respond not only to internal direction but to the large-scale structure of the environment. They sense where they are in a larger "field" of some kind and respond appropriately. For a moth, the field is a field of light emanating from, say, a light bulb. For a torpedo, the field is a sound field, radiating from a target ship. For a cell in a developing organism, the field might be a chemical gradient within the embryo, or a gene-activation region, or even a physical object like a membrane. For an entity navigating by GPS, it is the electromagnetic signal field coming from a satellite.

I use the phrase "upper direction" for this property of teleological entities, their taking direction from a larger field that contains them (McShea, 2012). The word "field" here is understood broadly to describe any large containing structure that acts causally on an entity within it. A gas molecule in a balloon receives upper direction from a larger entity consisting of the plastic of the balloon plus the other gas molecules within it. The balloon plus the other gas molecules constitute the "field." The contrast is with what I call "lateral direction." A laterally directed entity is acted upon only by entities of about the same scale that do not contain it. Imagine a long line of dominos standing upright on a table, spaced to cascade, so that each will knock down the one behind it when knocked over by the one ahead of it. The causal effect of each domino on the one behind is lateral direction. Generally speaking, causation in serial chains is lateral. The execution of the instructions by a laptop computer is mainly lateral, one flip-flop triggering another downstream. More abstractly, the heart of a Turing machine—a tape coding a series of instructions read sequentially—is lateral causation. But parallel chains can also be lateral. The interactions of gas molecules in an unconfined freely expanding gas occur laterally in parallel, with many interactions among molecules taking place at once. In a freely expanding gas, unlike a balloon, there is no larger structure that constrains a contained molecule, no containing structure that acts causally on it (or at least, containing structures are sufficiently large and distant that their effects can be ignored). Importantly, the distinction between downward and lateral causation is not black and white. Cases can be imagined that are partly both.

In any case, the proposal here is that teleology requires upper direction. I think this is a fairly intuitive view. At least, it is hard to imagine how persistence and plasticity could be achieved without it. Under purely lateral direction, with no larger organizing field, how is a teleological entity going to correct for errors, or to navigate from alternative starting points?

2.3. A counterexample?

Consider a case of goal-directedness, of persistent and plasticity, where upper direction might seem not to be necessary. Imagine a Lego robot that is directed to an X on the floor by a set of if-then instructions governing its lateral interactions with objects in the room. For example, suppose the instructions say that if it detects the couch, then it should turn 20 degrees to the right and move forward 1.2 meters. Then when it sees the TV, it is directed to turn 15 degrees to the left, and proceed forward 0.6 meters. And so on. Further suppose that there are a huge number of these instructions, enough to cover all contingencies, enough to set the robot on a trajectory toward the target from anywhere in the room, from anywhere that errors in navigation or movement could take it. In this setup, it might seem that upper-direction has been circumvented, that the robot will be able to find the target persistently and plastically, based only on the lateral direction it receives from the environment its (extensive list of) internal instructions. But that is not the case. In fact, such a robot is relying on the large-scale structure of its environment. That is, it relies not just on existence of particular objects in the room, on the existence of the couch and the TV and so on, but also on the spatial relationships among these objects. To see this, imagine what would happen if—in the middle of the robot's run—the relative positions of the objects were changed. Obviously, those spatial relationships are crucial to producing the robot's goal-directed behavior. In effect, they constitute a field, a structured object field, which is providing upper direction to the robot. And the robot is able to show persistence and plasticity owing (partly) to its immersion in it.

What makes this example instructive is that the setup draws so much attention to the internally programmed list of instructions that it is tempting to think of those instructions as the locus of teleology, as sufficient for the goal-seeking behavior of the robot, ignoring the crucial contribution of the "object field" that it is

immersed in. In thinking about goal-seeking behavior in biology, we tend to make the same mistake. For example, in explaining how a microorganism follows a chemical gradient, we tend to focus on the complex signal transduction mechanisms within the microorganism, or the gene-switching cascade that gives rise to those mechanisms, overlooking the enormously important contribution of the gradient itself, the field.

2.4. Other requirements

So teleology requires upper direction. But not too much upper direction. A ball rolling downhill through a narrow tube reliably arrives at the opening at the bottom, run after run. And this behavior does not strike us as teleological. For an entity to show persistence, it needs to have enough independence to deviate occasionally. Obviously the ball in the tube lacks the capacity to correct for errors. But also missing is the ability to *make* errors, the capacity for deviation. Without error, without some degree of independence, there can be no persistence, and the entity's behavior will not look teleological. Thus, in addition to upper direction, teleology requires independence. More precisely, the teleological entity must be partly upper directed and partly independent. This requirement for some degree of entity independence rules most machines out of the realm of the teleological. Machines are famous for their stereotypical behavior, in other words, for their reliability, for their *inability to deviate*. Interestingly, most machines are upper directed, in a sense, albeit not in a very interesting way. The blender or electric toothbrush that I plug into a wall socket is driven by a field, an electric field, created by the voltage differential across the prongs of the plug. But its behavior does not look goal directed because, like the ball rolling through a tube, it always does exactly the same thing. The same goes for the booting up of a laptop. Of course, once booted up, the laptop does have alternative behaviors, but they are all the direct result of operator inputs. Given a set of inputs, the behavior is completely, reliably stereotypical. And without any behavioral independence, it does not look teleological.

There is another requirement. The upper direction must be stable, at least relative to the behavior of the teleological entity. A torpedo following a sound field from a target ship will not seem to behave teleologically if the sound field is varying erratically (due, say, to a passing pod of whales). A Lego robot guided by an onboard GPS will not behave teleologically if the electromagnetic field from the satellite is fluctuating wildly (due, say, to interference from local cell phone transmissions). A moth will not behave teleologically if the light source it is trying to approach is changing rapidly (if say, the light is being turned on and off, or if the light is a flashlight, if the beam is being waved about). The field need not be totally invariant. But persistence of the teleological entity requires that the field be fairly stable on the timescale at which the entity makes errors and corrects for them.

Finally, stable upper direction is not quite enough. The gas molecule in the balloon is upper directed, but we don't think of it as teleological. We do not see an entity's behavior as teleological when we can clearly see and understand the upper-directed, error-correction mechanism at work. A ball rolling around in a bowl, driven toward the bottom by a gravitational field, might be said to "seek" the bottom, but the mechanism is simple and obvious and so the behavior does not seem teleological. In my earlier paper, I argue that we need one more thing, complexity. A certain amount of complexity, even mystery, must be present before we are moved to think about a system in teleological terms (see McShea, 2012).

2.5. Virtual nestedness: thermostats and DNA

In some teleological systems, the nestedness or containment required for teleology appears to be only virtual. Consider a thermostat that controls the temperature of a house. Let us say that when the air temperature rises above the set point, the thermostat activates the house's cooling system. And when the temperature falls below the set point, it activates the heating system. Now clearly the teleological entity is not the thermostat itself. What is behaving teleologically is the air in the house, more precisely, its temperature. And the space in which the temperature is changing is not a physical space but a phase space, a single axis representing temperature, with the temperature of the house represented by a point on that axis (Fig. 1). The point behaves teleologically, showing persistence and plasticity, moving back toward the set point whenever a fluctuation occurs. We could represent the "field" producing persistence and plasticity as a vector field, as in Fig. 1. There is containment here—a point contained within a vector field—but it is virtual, not physical. However, notice that in real thermostat-controlled systems, physical containment is present anyway. The control of the air temperature works better if the air is contained by the walls of the house. Without the walls, and the insulation they provide, the air temperature *could* be made to behave teleologically. It would be possible to regulate the air temperate of a wall-less house using a tremendously powerful heating and cooling system. But the system works much better with walls. Physical containment seems to be an effective and perhaps even crucial aspect of virtual containment.

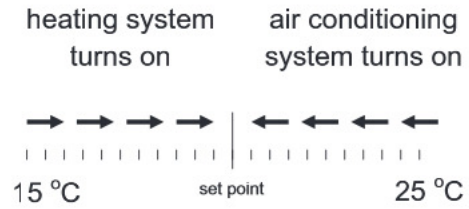

Fig. 1. Phase space for air temperature in a thermostat-controlled house, showing the vector field controlling temperature. (See text.)

Or consider a cell, perhaps a free-living amoeba. The concentrations of many of the molecular species within the cell seem to behave teleologically, showing persistence and plasticity in the face of perturbations. Let us understand this teleological behavior to be controlled by the DNA within the cell. For some molecular species dissolved in the cell's cytoplasm, suppose that its concentration is regulated by a few genes in the nucleus of the cell. As in the thermostat case, the concentration can be understood as a point contained within a vector field driving it back to some set point following deviations. The containment is virtual, not physical, again as in the thermostat case, and notice that here too a key to the effectiveness of the system's regulatory controls is containment. Containment within the cell's membrane prevents changes in concentration in the cytoplasm from dissipating before they reach the nucleus. In other words, the membrane forces changes to propagate back to the nucleus, producing the necessary feedback. Physical containment is not required in principle, but it turns out to be a convenient physical way to produce a virtual system with the right properties.

3. Animal wanting

Suppose that the thinking in the last section is right. What can we say about animal behavior driven by wanting-preferring-caring? Before answering that question, I need to say something about what is *not* being asked here, because much of the literature in psychology in this area has concerns that are beside the point. Wanting, preferring, and caring are understood here as types of (calm) emotion, but the concern here is not with these emotions

in their role of appraisal (Scherer, Shorr, & Johnstone, 2001), about their embodiment (Damasio, 1998), or about their role in moral judgment (Haidt, 2006). Further, the concern is not with the triggers of any given want, preference, or care. It is not with how seeing the cheesecake in the bakery window makes me want it. Nor is the concern with the regulation of wanting, with how I control or fail to control my desire for the cheesecake. It is not with the various rewards, punishments, and other inputs and biases that affect choices, with how the glucose content of my bloodstream or how the warnings of my doctor affect whether I will ultimately buy and eat the cheesecake. Rather, the concern is with the causal role of wants, preferences, and cares, with what they *are* and how they cause behavior.

If the argument of the last section is right, we can say that because wanting-caring-preferring is teleological, it must involve upper direction, which in turn requires containment, physical or virtual. And if virtual, physical containment is still likely to be involved. A want is a kind of field, like an electric field, and a behavior is like a charged particle moving within the field. The field powers the movement of the particle and gives it a general direction, without determining the details of its movement. The analogy is imperfect. For one thing, behavior-generating mechanisms have an internal structure, which allows their response to a driving field to be quite complicated. For another, at least in humans and probably in most mammals, wants are many. And each is probably multidimensional, with even simple wants underlain perhaps by more than one independent field. Further, even simple behaviors may be driven by more than one want. A mouse in a maze may want to run—after days in a cramped cage—at the same time as it wants the food that it smells. My decision to go to the supermarket and the act of going there may be the result of more than just my wanting food. It may also be wanting to get outdoors for a while. And wanting to get a feeling of accomplishment in an otherwise wasted day. And wanting to elevate my social status by owning and being seen to own some special brand of some food. And so on.

Further, for animals like us, especially mammals, it is clear that behavior mechanisms themselves are hugely complex, offering a space of behavioral options that is immense, and expanded orders of magnitude beyond immense by its combinatorics. Behavior is motor activity, and every motor activity can be undertaken in innumerable different ways. I don't only have the option of going to the store or not. I can walk. I can drive. I can go now. Or later. I can take a long route that satisfies some other want. Whatever route I take, long or short, there are many possible sequences of left and right turns, each a distinct motor sequence, a distinct behavior. I can put on my best clothes before going. Or not. I can talk on the phone along the way. All of these are behavioral options satisfying the want(s) that corresponds to "wanting to go to the store."

Still, despite the limitations of the particle-field analogy it is worth pointing out that it is broadly consistent with standard thinking in psychology about how motivation works (Berridge, 2004). Motivation has been taken to be a homeostatic mechanism, restoring the organism to equilibrial setpoints (e.g., hunger, driving the organism toward satiety). And the particle-field analogy can also be understood that way, as homeostatic, to some extent. Homeostasis is certainly teleological in the same sense (although some wanting—like my wanting to win a point in tennis—does not seem to be homeostatic in the sense of equilibrium restoring). Drive theory is also broadly consistent with the present view. Drives not only cause behavior but they do so in a way that permits other variables to intervene. In other words, drive theory allows for the possibility that drives are not fully determinative, that behavior can vary independently to some extent, just as a particle in a field can move independently to some extent.

What must wanting fields in humans be like? How are they constructed? The first answer is that we do not know. Human behavior could be driven like a torpedo in a sound field or a moth in a light field, with behavior-generating mechanisms physically contained within a physical field of some kind. The physical field would be the physical manifestation of wanting-preferring-caring. In principle, this could be a local electrical, neurochemical field, or gene-expression field lying entirely within some discrete brain structure. And indeed certain brain structures have been identified as involved in motivation. For example, the mesolimbic system is thought to be involved in nonconscious wanting. However, conscious wanting seems to involve other brain structures (Berridge, 2004), suggesting a more globally distributed mechanism. Likewise, in principle, it could be that the behavior-generating mechanism lies within some discrete neural organ, as smiling seems to lie within the motor neocortex. However, voluntary behavior generally is more distributed. Still, even if behavior generation is somewhat distributed, it could be that wanting-preferring-caring is even more widely distributed, among brain regions and perhaps even embedded in the network connecting brain regions (Symmonds & Dolan, 2012). Could simple physical containment in a near-global physical field still be a real possibility?

Alternatively, teleological behavior could involve virtual containment, like a thermostat, a virtual point, representing behavior, contained within a field representing wanting-preferring-caring, with physical containment of some sort to enforce the feedback required for persistence and plasticity. What about the possibility that there is no physical containment at all? In that case, we would want to ask how persistence and plasticity are achieved. Real systems like thermostats and cells ultimately depend on physical containment to achieve virtual containment. If the brain does it in some other way, then how?

A remark is needed on the common notion of a motivational hierarchy in the brain consisting of lower and higher cortical centers, with higher forebrain structures controlling the lower, more distal and primitive brainstem structures. The wiring among these structures is known to be sufficiently complex as to raise doubts about the existence of any sort of hierarchy among them, but in any case it is clear that this notion is invoking hierarchy in a very different sense than in the particle-field analogy. The motivational hierarchy is thought to be a command hierarchy but not a physical hierarchy of containment. An army is just such a command hierarchy, with colonels giving commands to majors and lower ranks, but it is not a hierarchy of containment. Majors are not contained within colonels. In the argument offered here, wanting is thought to direct behavior but to do so by virtue of containing the behavior-generating mechanisms, by virtue of physical envelopment.

Also, it may seem that something crucial is missing from the picture of wanting-preferring-caring developed here, namely consciousness. The concern in this essay is with *deliberate* thought, speech, and action, and therefore some degree of consciousness is clearly necessary. Not necessarily self-consciousness, or a view of oneself as a thinking being with a life extended in time, but simple awareness, of the sort that both mouse and human clearly have. So it is clear that consciousness is necessary for deliberate behavior but it is also pretty clear that like logic, it has no motive force. The conscious mind observes the world, and takes notice of the organism's wants, preferences, and cares, but consciousness by itself can do little more than observe and notice. True, if there is an affective response to what is passively observed and noticed—a goal directed, valenced, seeking, or purposeful response—it may be triggered by the contents of consciousness. But the machinery that is triggered, that does the responding, is the wanting-preferencing-caring mechanism, not consciousness.

Interestingly, the standard folk psychology of wanting is consistent with the above view. We think of wants as long-lived, relative to the behaviors they motivate. I want to go the store more or less continuously, even if somewhat in the background, the whole time

I am thinking about alternative ways to get there. We think of wants as stable, as difficult to change by thinking or acts of will. I wish I didn't care so much about the upcoming election, but I do. We think of behavior as flexibly independent, many alternative behaviors being consistent with a single want or set of wants. There are lots of ways to make a good impression. We think of behavior as subject to deviation, to momentary whims. I want to lose weight but found myself eating that third piece of cheesecake anyway. It is even subject to unmotivated deviation. En route to meeting a friend, I slip on the sidewalk and fall down. But behavior is also plastic and persistent. I really do want to be healthy, and so behaviorally I return to my low fat diet the next day after eating the cheesecake. After slipping, I get up off the sidewalk and continue on my way to meeting the friend. Now wanting fields seem to be numerous and complexly related to each other. And the behavior particles they drive doubtless have a complex internal structure. But these complexities aside, consistent with folk psychology, the relationship between wanting and behavior does seem to have the essential features of a field-particle relationship.

4. Machine wanting

4.1. Logic machines

Taking Hume's point seriously, taking the logic-wanting gap as a deep truth about the nature of mind, what follows? It follows that no exercise in logic is going to produce wanting. To the extent that our machines are logic machines, they are not going to want anything, prefer anything, care about anything. Suitably programmed, my laptop is an excellent logic machine. But imagine it lying in the middle of the road (where I am often tempted to put it), with a bus bearing down it. The lid of the laptop is raised and its camera eye is trained on the oncoming bus. Consistent with Hume, it accepts the situation with the same indifference as does the pebble lying next to it. If it could put words to its response to the oncoming bus, it might be a somewhat bored, "So what." Well, that is misleading. Boredom is an affective state. A laptop cannot even be bored.

Logic machines cannot want, prefer, or care. It follows that they cannot deliberately do or say anything either. All deliberate behavior in humans is, in Hume's terms, volitional, that is, driven by wants, preferences, cares. Mr. Data of the old Star Trek series was supposed to be a purely logical being. Mr. Spock of the even-older series struggled to be one. But a purely logical being would have no motivation to get up in the morning, much less to interact with his starshipmates or to do his job. He could, obviously, be programmed to perform the right behavior, to get up, follow orders, interact with others, etc. But as a being of pure logic, he could not want to do them. No want follows from any chain of logic. Mr. Data and Mr. Spock are not purely logical beings, nor could they be.

Suppose then that I confront a pure-logic machine, and I ask it: "What would you like to do this afternoon?" And suppose it responds. What should we make of that response? The answer is that if the machine's response is generated by logic, then it cannot be a real answer to the question. Suppose we program a machine to respond to the question using the following algorithm: Find out what month it is and find out whether or not the back 40 have been plowed yet (consult memory or ask someone). If it is April and if the back 40 have not yet been plowed this year, then answer the question by saying you would like to plow the back 40. If it is not April or if the back 40 have not yet been plowed, remain silent. Then, goto step 2 to consider the possibility of answering that you would like to clean the house. The machine marches through a series of such steps, eventually uttering the statement: "I would like to walk the dog." The procedure seemed logical, and the result was an apparent statement of preference. However, if Hume is right, a preference must be a physical state or an event, so the question arises: which physical state or event, or set of states or events, corresponds to a preference? We have three options. We could say that the machine's physical state(s) as it performs the if-then operations in the sequence above just is a preference. But this is an uncomfortable conclusion, because none of those if-then statements seems at all like a preference. Preference statements in ordinary thought and speech require no if-thens, explicitly or implicitly. (I can say that if it rains, I will want to run for cover, but it's easy to restate the same want—I want to stay dry—without any conditional.) But we would be forced to say that one or all of those if-thens just is a preference, even though it bears no resemblance to one. A second option is that we could say that the machine's physical state as it utters the concluding statement, "I would like to walk the dog," just is the preference. But this too leads to an uncomfortable conclusion. A machine programmed to do nothing but utter the statement, "I would like to walk the dog," could be in the same physical state as the one above, and therefore would also have to be given the same credit for having a preference. And that too sounds wrong. The third alternative is that the machine only seemed to answer the question, that it had no preference, and that therefore its "answer" to the question is not an answer at all.

4.2. Machines that want

Can we design a machine with some degree of actual or virtual upper direction, capable of real goal directedness? In principle, I do not see why not. Modern computers are mostly serial processors, rich in lateral direction and poor in upper direction, with hardware that operates like falling dominos or Lego robots. On the other hand, there is no reason in principle that we could not build a different sort of machine, one with a physically hierarchical design, consisting of a behavior-generating mechanism nested within physical fields of some kind. Alternatively, we could take a virtual approach, using laterally-directed hardware to simulate a virtual complex upper-directed system.

What are the requirements for such a design? The argument above suggests that for teleological behavior—for persistence and plasticity—we need a behavior-generating mechanism that is upper directed but still partly independent. And the upper-level field directing the behavior needs to be relatively stable. But these are minimal requirements. Real human wanting is more complicated, of course. For one thing, it is multidimensional. We want many things at the same time. At this moment, I want to finish this paragraph in time to rush off to a meeting. I foolishly skipped breakfast, so I also want something to eat. I want to phone my mother, who I have not spoken to in some days. And then on a longer timescale, I want to improve my tennis game. I want to plan a family adventure. And on an even longer timescale, I want my family to be happy. I want to stay healthy. I want to spend more time on my nonacademic writing. And so on. My wants are not only different on different timescales, but they vary in specificity. I want to finish *this* paragraph, and I want to do it right away. But I do not know where I want to take my family, or what sort of adventure I want it to be. The want associated with the adventure is still poorly specified, unlike the want associated with finishing this paragraph. Also, wants are weighted. It turns out that I want to work on this paragraph more than I want to be on time for that meeting (a conclusion supported by the fact that the meeting has now started and I am still sitting here writing). Interestingly, the want associated with the meeting, although overridden, still hovers, distracting me from the writing. In other words, my wants are multiple and simultaneously present, and they interact. Finally, wants are developmental. My short-time-

scale wants change from moment to moment. My long-timescale wants change over decades. And all of my wants fluctuate in degree of specificity. My fairly unspecific wanting to plan a family adventure will (I hope) crystalize into a very specific plan, involving particular destinations, dates, etc. In sum, a realistic simulation of human wanting would not only be higher-level-field based (to meet the requirements of teleology), but it would involve many higher-level fields simultaneously, and those fields would be weighted, interactive, varying in degree of specificity, and perhaps—depending on the degree of realism sought—developmental.

It is often said that animals generally, and humans in particular, are motivated only by survival and reproductive success. It is often said, but it is transparently false. It is true that our motivational structure is (partly) the product of millions of years of natural selection favoring just these two results. However, survival and reproduction are ultimate goals, the causes of the design of our brains, just as they are the causes of the design of a bird wing. But proximately, bird wings are for flight, not survival. Likewise, our proximate goals, our motivations, usually have little to do with survival or reproduction. Proximately, humans are motivated by general affective states like empathy, greed, lust, love, joy, social approval, envy, and so on, and by the more-specific wants that flow from them, and only rarely by survival and reproduction. Whatever their evolutionary history, our affective states and our wants are for us ends in themselves. Obviously, our behavior would be a lot simpler, more stereotypical, more "robotic," if they were not, if we only had two motivations, survival and reproduction. To put it another way, for most readers, your own survival and reproduction probably never crossed your mind in the course of reading this essay. So why are you reading it?

4.3. Physical versus virtual

Shall we try to simulate human wanting with a physical field or a virtual one? There is some reason to be leery of the virtual approach. A simulation of a thermostat-plus-heating-cooling-system may capture most of the dynamics of the real thing, but it will not warm or cool a house. And without the real-world test of success—the actual warming or cooling of a real house—there is no way to know whether we have captured all of the central features of the process being simulated. For example, if we did not know anything about warming a house beyond what is captured by the vector field in Fig. 1, if we had no experience with or did not understand real heating-cooling systems and real houses, it might never occur to us that real houses need walls (not only to keep the warm and cold air inside, but—say—to protect the system from damage from outside). And walls have consequences for the dynamics of the system. By the same reasoning, without knowing how wanting really works in people, there is a danger that a purely virtual simulation could miss something crucial. Physical systems, unlike abstractions, have physical constraints, the occasional unforeseen property that turns out to be crucial to the success of the simulation. There is so much mystery around wanting that it might make sense to attempt our first simulations in physical systems with lots of physical constraints. At worst, our failures might start to reveal what sort of physical systems will not work, and that might help to aim us in the right direction. We are shooting almost in the dark here. Every clue could be helpful.

5. A thin account

Hume is right. Wants, preference, cares cannot be rational in the formal sense of the word. They do not follow logically from any set of facts, statements, or observations about the world, nor can they be contradicted by any set of facts, statements, or observations. Wants are states or events. They are "original existences." They are something physical that happens. Not statements or propositions that follow logically.

My argument about the hierarchy requirements for teleology may or may not be right. The idea is too new and has not yet been sufficiently vetted. Still, if it is right, it tells us something that could be potentially useful in the quest to simulate human wanting. Admittedly it tells us only a little something. Nothing about the sensory or thinking mechanisms that implement behavior. Nothing about the role of consciousness and cognition in triggering wanting. It tells us that teleology requires upper direction and probably physical containment, that wanting must consist of a behavior mechanism (partly) contained within and directed by a larger field. And in that case, if we want to build a machine that wants, this view suggests that laterally structured computational designs are not going to work. We are probably going to need a new kind of kind of hardware.

If we finally succeed, if we eventually figure out how to build a machine that can honestly tell us what it would like to do this afternoon, what will it say? For most questions about wanting-preferring-caring, the answer will, I would think, be highly predictable. We need only calculate the resultant of the various fields directing its behavior. And since we will have chosen those fields, and the various inputs that affect their intensities, it should not be hard to calculate that resultant, at least to calculate the expected, on-average resultant. Of course, if the machine wants to win the card game it is playing, and also wants to have something to eat, and also wants to open the window to let in more of whatever is the machine equivalent of fresh air, and also wants to answer the ringing telephone, and so on—over a long list of simultaneous human-like wants of varying specificity on varying timescales—even the expected the vector sum may be difficult to calculate. (Especially if the fields and their interactions are complicated.) Still, how it chooses will depend on known, or at least knowable, factors, the structure and intensity of the fields we have chosen to instantiate those wants.

Still, notice that as wanting questions go, our question to the machine about its preferences for this afternoon is a little unusual. It seems to suggest a situation with none of the urgency of a card game or a ringing telephone, a situation in which strong wants are absent, in which many weak wants are balanced on a knife edge, and in which the accident of their present intensities will be decisive. The whims of the moment. The decision could be made in some very-difficult-to-accurately-calculate tenth decimal place. At the same time, the question seems to invite the machine to bring to bear on the decision the full range of wants, preferences, and cares of which it is capable. If going for a walk alone in the woods is an option, it might consider not only the momentary desire for fresh air, but also the long-run desire to lead a healthier life, assuming walking is healthy for this machine and that health is desired. Our wanting-caring-preferring machine might even choose this option in response to an ever-present desire for solitude, the kind of solitude that allows it to reflect calmly on its life, to examine the full range of it wants and the various ways it might satisfy them. Or it might make this choice out of a desire to satisfy a deep aesthetic arising from movement, from making one's way in the world, from envelopment in nature. Maybe all of these would be in play. Maybe none. Depends on how we build it. Depends on what it wants.

Acknowledgments

This paper was inspired by a conversation many years ago with the brilliant painter, teacher, and polymath Louis Finkelstein, who once posed for me the focal question here: What would it take for a machine intelligence to answer the question, "What would you like

to do this afternoon?" This is, at the moment, the best answer I can give. Also, for some very useful comments on the manuscript, I thank Jessica Huang, Sankar Kannusamy, Anne Talkington, Stan Salthe, and an anonymous reviewer.

References

Ariely, D. (2010). *Predictably irrational*. New York: Harper Perennial.
Berridge, K. C. (2004). Motivation concepts in behavior neuroscience. *Physiology & Behavior, 81*, 170–209.
Damasio, A. R. (1998). *Descartes' error*. New York: Bard/Avon Books.
Darwin, C. (1872 [1998]). *The expression of emotion in man and animals*. New York: Oxford University Press.
Dolan, R., & Sharot, T. (Eds.). (2012). *Neuroscience of preferences and choice*. Amsterdam: Elsevier.
Haidt, J. (2006). *The happiness hypothesis*. New York: Basic Books.
Hume, D. (1740 [1978]). In L. A. Selby-Bigge (Ed.), *A treatise of human nature* (2nd ed.). Oxford: Oxford University Press [1978].
James, W. (1890). *The principles of psychology*. New York: Holt.
McShea, D. W. (2012). Upper-directed systems: A new approach to teleology in biology. *Biology and Philosophy, 27*, 663–684.
Nagel, E. (1979). *Teleology revisited and other essays in the philosophy and history of science*. New York: Columbia University Press.
Salthe, S. N. (1985). *Evolving hierarchical systems*. New York: Columbia University Press.
Salthe, S. N. (2009). A hierarchical framework for levels of reality: Understanding through representation. *Axiomathes, 19*, 87–99.
Scherer, K. R., Shorr, A., & Johnstone, T. (Eds.). (2001). *Appraisal processes in emotion: Theory, methods, research*. Canary, NC: Oxford University Press.
Simon, H. A. (1962). The architecture of complexity. *Proceedings of the American Philosophical Society, 106*, 467–482.
Simon, H. A. (1967). Emotional and motivation controls of cognition. *Psychological Review, 74*, 29–39.
Symmonds, M., & Dolan, R. J. (2012). The neurobiology of preferences. In R. Dolan & T. Sharot (Eds.), *Neuroscience of preferences and choice* (pp. 3–31). Amsterdam: Elsevier.
Wimsatt, W. C. (1974). Complexity and organization. In K. F. Schaffner & R. S. Cohen (Eds.), *Philosophy of science association 1972* (pp. 67–86). Dordrecht, Netherlands: D. Reidel.
Wimsatt, W. C. (1976). Reductionism, levels of organization, and the mind-body problem. In G. G. Globus (Ed.), *Consciousness and the brain*. Plenum Press.
Wimsatt, W. C. (1994). The ontology of complex systems: Levels of organization, perspectives, and causal thickets. *Canadian Journal of Philosophy, 20*(Suppl.), 207–274.