Reconciling Mechanistic and Non-Mechanistic Explanation in Cognitive Science

William York[1]
[1] Indiana University, Bloomington, IN U.S.A.
wwyork@indiana.edu

Abstract

Mechanistic explanation plays a central role in cognitive science. However, there are phenomena of interest to cognitive scientists for which mechanistic explanations are either not feasible or not appropriate—at least, not on their own. Examples would include any phenomena in which the "system" under consideration includes autonomous agents (e.g., human beings) among the entities that make it up. This is the case with humanistic phenomena such as aesthetic or artistic experience, among other examples. In such cases, a satisfying account would have to take into account the role of the individual person in relation to a surrounding cultural, social, and historical context. In this paper, I argue that non-mechanistic explanations must be given a legitimate place in cognitive science if the field wants to contribute usefully to our understanding of such phenomena. However, I also stress that recognizing the value of non-mechanistic explanations does not equate to being anti-scientific or even anti-mechanistic. Finally, I offer some suggestions for how mechanistic and non-mechanistic accounts might complement one another in practice. I do so by focusing on the phenomenon of aesthetic sensibility.

1 Introduction

In cognitive science, explanations are most often given in terms of mechanisms (Wright & Bechtel, 2007). These mechanisms can be computational, neural, diagrammatic, or even physical (e.g., in robotics). However, there are phenomena for which mechanistic explanations are either not feasible or not appropriate. Examples would include any phenomena in which the "system" under consideration includes autonomous agents (e.g., human beings) among the entities that make it up.

Cognitive science and AI have been subject to various anti-mechanist critiques over the years. Such critiques have often been met with derision and dismissed as mystical or anti-scientific. However, the more thoughtful of these anti-mechanist critiques (Shanker, 1998; Dreyfus, 1992; Shanon, 2008) raise issues that are worthy of careful consideration, even for the staunch mechanist. To paraphrase Shanon, mechanistic explanation in cognitive science is a means to an end—that of understanding the mind—not an end in itself. Such challenges to mechanistic explanation's supremacy offer a potential corrective to overzealous or "greedy" reductionism (Wimsatt, 2007).

My aim in this paper is to reconcile certain aspects of the mechanist and anti-mechanistic viewpoints. I argue that non-mechanistic explanations must be given a legitimate place in cognitive science if the field wants to contribute usefully to our understanding of humanistic phenomena such as creativity, artistic experience, and aesthetic judgment. However, I also stress that recognizing the value of non-mechanistic explanations does not equate to being anti-scientific or even anti-mechanistic. Rather, there are phenomena that lend themselves to the sort of decomposition-into-parts

that mechanistic explanation offers, while there are other phenomena for which such an approach is inappropriate. Getting clear about levels of description is crucial to understanding why this is the case. Specifically, a holistic or non-mechanistic account given at one level of description can coexist with—and complement—a mechanistic account given at another (lower) level. Toward the end of the paper, I offer some suggestions for how mechanistic and non-mechanistic accounts might complement one another in practice, focusing on the topic of aesthetic sensibility.

2 Mechanistic Explanation

In recent decades, philosophers of science have effectively debunked the once-standard notion of scientific explanation as involving subsumption under laws—the deductive–nomological model (Hempel, 1998)—at least as it pertains to psychology and cognitive science. For one thing, there is a paucity of laws in psychology. Furthermore, what laws do exist (e.g., the Law of Effect) typically do not provide explanations; rather, they capture patterns or regularities that themselves stand in need of explanation (Cummins, 2000).

Rather than laws, then, it is mechanisms that play the primary explanatory role in these fields. According to Bechtel and Abrahamsen (2005), "A mechanism is a structure performing a function in virtue of its components parts, component operations, and their organization. The orchestrated functioning of the mechanism is responsible for one or more phenomena" (p. 423). The authors distinguish physical mechanisms (e.g., the Krebs cycle in biology) from mechanistic explanations. The latter are "epistemic products" and may not map directly onto physical processes involving "thing-like" entities. In cognitive science, mechanistic models can take many forms, from computer models to flow charts or box-and-arrow diagrams. Such models typically qualify as epistemic products in the above sense, although there are exceptions—for example, in cognitive neuroscience, where a mechanistic explanation might consist in relating an actual neurological mechanism to a particular cognitive function.

2.1 Why Mechanisms?

The value of mechanistic explanation is familiar to anyone who has ever peered under the hood of a car or taken apart a malfunctioning device in hopes of repairing it. The idea is that in order to understand how something works, the best approach is often to take it apart and see how the pieces fit together; or, conversely, to take a bunch of parts and try to build the thing out of them. This "understanding by building" ethos is what links cognitive science with AI and robotics (Ekbia, 2008). Again, the component parts need not be physical parts; they can just as well be abstract entities, such as the nodes and connections in a neural network.

While cognitive science obviously faces many unsolved problems, many of its successes—and many of its more enlightening failures—are due to this mechanistic approach. If you can build (or program) a system that accomplishes some task, then that system can be said to explain the underlying cognitive capacity being modeled. As Herbert Simon (1995) famously put it, "The moment of truth is a running program" (p. 96). However, this "running program" need not actually be a "true" explanation in the sense of mirroring the precise workings of the brain; even highly artificial or "toy" models can play an explanatory role. For one thing, they can serve as proofs of concept by showing that a given mechanism is *capable* of producing a certain kind of behavior (e.g., learning, categorization, etc.). At the same time, these models can shed light on the boundary conditions beyond which a certain kind of mechanism "breaks down."

2.2 Limits of Mechanistic Explanation

Despite its many successes, mechanistic explanation is not all-encompassing. But in what types of cases might we expect mechanistic accounts to falter? Here is one suggestion: If the "system" under consideration includes *autonomous agents*—for example, human beings as opposed to parts of human beings or processes that occur inside human beings—then purely mechanistic accounts are unlikely to succeed. Bechtel and Abrahamsen (2007) note, "Mechanisms, insofar as they involve purely causal processes, are fully determined in their responses and so lack the requisites for [personal] agency" (p. 96). They add that in order to account for human-level phenomena such as moral agency, "[W]e will have to move beyond the common conception of mechanisms as purely reactive systems responding only when confronted with a stimulus" (ibid.). Yet it is unclear how this move could be made without stretching the concept of *mechanism* beyond its breaking point.

3 Non-Mechanistic Explanation?

It is one thing to suggest that mechanistic explanation has certain limitations. It is another thing to suggest that mechanistic explanation is the wrong approach for *any* account of cognition or perception. Since their inception, AI and cognitive science have faced objections about whether human thought or behavior could possibly be explained mechanistically. Since mechanistic explanations in these fields often take the form of computer programs, these objections are closely related to the claim that human-like behavior cannot be explained in terms of computer programs: Whereas human behavior is flexible and unpredictable, computer programs are rule-governed and deterministic. Or so the argument goes.

3.1 Turing and the Argument from Informality of Behavior

Turing (1950) anticipated this sort of objection in his reply to the Argument from Informality of Behavior. In responding to this objection, Turing drew a distinction between *rules of conduct* (e.g., "Stop if you see a red light") and *laws of behavior* (e.g., "If you pinch someone, s/he will squeak"). He argued that even if there aren't rules of conduct for every situation, there could still be underlying laws of behavior at work—suggesting that humans could, in the proper light, be conceived of as machines. As such, he saw no reason why a suitably programmed computer could not achieve human-like intelligence simply in virtue of its being a machine.

Turing's conclusion is probably correct, though perhaps not for the right reasons, as I will discuss in Section 4. In the meantime, I turn to a couple of (relatively) recent critiques of mechanistic psychology—the first due to the philosopher Hubert Dreyfus and his brother, Stuart; and the second due to the psychologist Benny Shanon.

3.2 Dreyfus (and Dreyfus)

In effect, the Dreyfus's critiques latch onto what is intuitively right about the Argument from Informality of Behavior while shedding the naïve-sounding confusions that Turing introduces into it. Of particular interest are the characterization of the four different types of intelligent activity presented in Dreyfus (1992); and the five-stage model of skill acquisition outlined in Dreyfus and Dreyfus (1988).

Dreyfus (1992) designates the four areas of intelligent activity as (I) *associationistic*, (II) *simple–formal*, (III) *complex–formal*, and (IV) *nonformal*. Of these four areas, only the first two involve activities that are tractable for classical AI programs. Associationistic activities include sorting lists and navigating mazes via trial and error, while simple–formal ones include mechanical theorem-

proving and computable games such as tic-tac-toe. In these areas, there exist algorithms *at the level of the tasks themselves*, which is not the case with the other two areas. In particular, nonformal activities—which include everything from open-structured games (e.g., Twenty Questions) games to non-mechanical language translation—resist characterization in terms of algorithms at the task level. For example, while there is an algorithm one can follow in order to play tic-tac-toe without ever losing, no such algorithm exists for, say, Twenty Questions. In other words, while the tasks in Areas I and II can be carried out *mechanically* according to an algorithm, the ones in Area IV cannot be. Meanwhile, tasks in Area III fall into a gray area: There typically exist heuristics for complex–formal tasks such as chess-playing or non-mechanical theorem-proving, but according to Dreyfus, the *application* of those heuristics calls on the sort of context-sensitive, non-mechanical intelligence characteristic of Area IV. For these reasons, Dreyfus held that tasks in Areas III and IV were precisely the kinds of things that "computers can't do," to paraphrase the title of his controversial book.

The five-stage model of skill development put forth by Dreyfus and Dreyfus (1988) also emphasizes the distinction between mechanical and non-mechanical (or holistic) modes of behavior. The gist of this account is that as an individual develops a given skill—such as driving a car, learning a musical instrument, or playing a sport—s/he passes through a series of qualitatively distinct stages. In progressing through these stages—*novice*, *advanced beginner*, *competent*, *proficient*, and *expert*—the individual moves from an atomistic, explicitly rule-governed mode of behavior to a holistic, "involved" mode in which no conscious rule-following takes place. Thus, the five-stage model is *non-mechanistic* in the sense that it proposes no mechanisms, but it still plays an explanatory role. For example, it explains why an individual's performance is likely to drop off when s/he consciously reasons about how to carry out a given action (e.g., shooting free throws in basketball or playing a difficult passage on the guitar) as opposed to "just doing it." ("Why did he miss those free throws so badly?" "He was thinking too much.")

Yet from a different perspective, one might be curious to understand what the associated mental mechanisms are at each stage, or perhaps how the various "phase transitions" take place as the individual progresses from one stage to the next. Thus, the fact that we can explain the individual's behavior in non-mechanistic terms does not preclude us from delving deeper and trying to uncover the lower-level mechanisms that *give rise to* the behavior.

3.3 Shanon: Beyond Procedural Psychology

Shanon's critique of mechanistic psychology centers on inadequacies with the representational–computational view of mind (i.e., symbolic AI). Shanon highlights the deeply rooted problems with meaning and reference that even advocates of the representational–computational theory have been forced to acknowledge (Fodor, 1980). Essentially, Shanon's argument is that (a) if mechanistic accounts presuppose fixed, context-free representations and computational processes involving them, and (b) these processes cannot account for meaning and reference, then (c) such accounts are bound to be inadequate.

However, Shanon's argument extends to approaches that posit mechanisms at lower levels of analysis, the prime example being connectionism. His argument here is less convincing, as it is based on the requirement that psychological explanation must be couched in "genuinely psychological"—or intrinsically meaningful—terms. Since the units in a connectionist network are (typically) not intrinsically meaningful, connectionism is unable to meet the above requirement. (This conclusion is problematic, however, for reasons I touch on in the next section.)

If mechanistic or procedural approaches are to be cast aside, then what is the alternative? Shanon advocates a "non-procedural" approach, one that would shed the emphasis on underlying mechanisms and instead concern itself "with what there is on the surface" (p. 329). He elaborates,

[P]sychological investigation will be concerned with structural constraints on cognitive expressions, the dynamics of their progression in time, the functions they serve, the contextual dependencies associated with them, the course of their ontogenesis and the history of their evolution in culture and societies. (ibid.)

Such a research program—even in broad outlines—will strike many cognitive scientists as unsatisfactory, even sterile. However, the suggestion that we pay attention to historical, societal, and cultural factors is a welcome one (see Section 5). Even so, it is possible to do so without abandoning mechanistic explanation altogether. As I emphasize in the next section, there are any number of levels at which we might wish to understand a given phenomenon. Some call for mechanistic explanations, whereas others do not.

4 Getting Clear about Levels

Recall Turing's reply to the Argument from Informality of Behavior, which turned on a distinction between rules of conduct and laws of behavior. Turing was surely right to distinguish between the different *levels* at which rules, laws, or (for our purposes) mechanisms may apply. However, precisely where computers fit into this two-fold distinction is less cut-and-dried than Turing suggests.

For some, computers can do nothing but follow rules: they are "rule-following beasts" (Hofstadter D. R., 1979). To others, computers are no more capable of following rules than they are of disobeying them, since rule-following is an inherently normative, or social, activity (Shanker 1998). The trouble arises when one inadvertently slides back and forth between these different senses of *rule.*

For example, Turing drew an analogy between a "human computer"—that is, a human being whose job involves performing calculations by following a sequence of unambiguous rules—and the read–write head of a digital computer. In doing so, he was likening a whole system (i.e., the "human computer") to a *part* of another system (the digital computer). Bechtel elaborates,

In this way, an activity performed by humans provided the model for operations occurring in their minds. The explanatory strategy is comparable to that of physiological chemists' invoking fermentations as intermediate processes in alcoholic fermentation. The component operations within the posited mechanism are of the same sort as the behaviors of the mechanism itself (Bechtel, 2005, p. 210).

The mistake that Bechtel describes here is a common one. Hofstadter (1985) criticizes the same sort of mistake regarding classical AI's efforts to "build rational thought ('cognition') out of smaller rational thoughts" (p. 643).

Ironically, this kind of mistake is made by parties on both sides of the mechanist–anti-mechanist fence. On the mechanist side, there is the insistence that since brains are physical entities which obey the laws of physics, the thoughts and behaviors they give rise to must be equally law-like. On the other side, we find those who, like Shanon, argue that since thoughts and behaviors are primarily *not* mechanical or law-like, it is misguided to investigate the mechanisms that give rise to them. What both sides seem to be missing is an appreciation for the complementary roles played by reduction and emergence in relating such levels to one another. That is, the entities at a lower level (e.g., statistical mechanics) are typically very different from the entities at a higher level (e.g., thermodynamics). This is what is puzzling about Shanon's requirement that the lower-level entities in a psychological explanation be "genuinely psychological." It is akin to arguing that the entities of statistical mechanics are not "genuinely thermodynamical."

5 Reconciling Mechanistic and Non-Mechanistic Explanation

Are there phenomena for which a non-mechanistic account at one level of abstraction can be seen as complementary to a mechanistic account at another (lower) level abstraction? My provisional answer is "yes." The key here, once again, is to distinguish between different levels—specifically, between the personal and the sub-personal levels. To see what this reconciliation might look like, I will focus on the phenomenon of aesthetic sensibility, for which a variety of mechanistic accounts have been offered, albeit with limited success.

5.1 Aesthetics: Personal vs. Sub-Personal Perspectives

In what we call the Arts a person who has judgment develops. (Wittgenstein, 1967, p. 5)

When cognitive scientists have turned their focus toward aesthetics and the arts, the emphasis has typically been on internal processing mechanisms (Leder, Belke, Oeberst, & Augustin, 2004; Reber, Schwarz, & Winkielman, 2004). But rather than just asking what is going on inside the mind/brain of a person who is making an aesthetic judgment at a given moment, we might instead step back and ask how it is than an individual develops a personal sense of taste—an aesthetic sensibility—over time. For example, how does a person go from initially hearing an unfamiliar genre of music as "noise" to developing an appreciation for the nuances of the genre? This process typically unfolds over a span of years, even decades. Surely, there are neurological changes that accompany this process, but a genuinely enlightening account of this progression would seemingly have to move beyond the realm of inner mental processes and account for other variables.

To put it differently, the relevant entities in this developmental process are not (just) at the *sub-personal* level of inner mental processes, but at the *personal* level. These "entities" include cultural artifacts (e.g., recordings, books about music, etc.), institutions (record stores, concert halls), and *other people*. That is, we often listen to or discuss music with others, and these experiences help shape our own sensibilities. As such, a strictly mechanistic account of mental processing that stays "inside the head" is likely to misconstrue the phenomenon—to claim too much on behalf of such inner processes while saying too little about other relevant factors. Here it seems likely that the sort of "non-procedural" explanation advocated by Shanon is likely to be more enlightening than—or at least a much-needed complement to—the computational "inner process" models that typically predominate when cognitive scientists examine aesthetic phenomena.

5.2 "How Do People Do It" vs. "How Does the Mind/Brain Do It?"

It has often been said that a person doesn't really understand something until he teaches it to someone else. Actually a person doesn't really understand something until he can teach it to a computer, i.e., express it as an algorithm (Knuth, 1973, p. 709)

If we look at aesthetic judgment in terms of Dreyfus's four areas of intelligent activity, it clearly fits into Area IV—the nonformal area. Knowing this helps to explain why we are largely unable to articulate the "steps" that give rise to a particular aesthetic judgment—namely, because we aren't "following" any such steps. Rather, when we seek an explanation for some aesthetic judgment in the context of everyday life, we are not seeking a mechanistic explanation. As Wittgenstein put it,

Suppose a poem sounded old-fashioned, what would be the criterion that you had found out what was old-fashioned in it. One criterion would be that when something was pointed out you were satisfied. And another criterion: "No-one would use that word today" (p. 20).

This is not a mechanistic explanation, nor would one be relevant here. Furthermore, since there are no explicit steps that a person *could* follow in making such a judgment (i.e., that a given poem sounded old-fashioned), one could not teach this capacity to a computer by feeding it a "poem judging" algorithm. However, contrary to Knuth's claim, it would not follow that we didn't *understand* what it means to make such a judgment. Rather, the understanding gained would amount to realizing that such judgments are inherently nonformal (or non-mechanistic) *at the personal level*—regardless of the underlying mental or computational mechanisms at the sub-personal level. [*]

6 Conclusion

Cognitive science faces unique methodological challenges due to the broad scope of its subject matter. In this paper, I have argued against a "one size fits all" approach to explanation. That is, there is space for mechanistic as well as non-mechanistic explanations. Whether or not humanistic phenomena such as aesthetic sensibility are reducible to lower-level mechanisms in principle, they are not reducible to such mechanisms in practice. A genuinely enlightening account of such phenomena requires that we remain mindful of the different levels of description involved, as well as the kinds of explanation that are appropriate at these different levels.

Bibliography

Bechtel, W. (2005). Mental mechanisms: What are the operations? . *Proceedings of the 27th Annual Meeting of the Cognitive Science Society*, (pp. 208-213).

Bechtel, W., & Abrahamsen, A. (2005). Explanation: A Mechanistic Alternative. *Studies in History and Philosophy of the Biological and Biomedical Sciences, 36*, 421-441.

Bechtel, W., & Abrahamsen, A. (2007). Mental mechanisms, autonomous systems, and moral agency. *Proceedings of the 29th Annual Cognitive Science Society* (pp. 95-100). Austin, Tex.: Cognitive Science Society.

Cummins, R. C. (2000). "How does it work" versus "what are the laws?": Two conceptions of psychological explanation. In F. Keil, & R. A. Wilson (Eds.), *Explanation and Cognition* (pp. 117-145). Cambridge, Mass.: MIT Press.

Dreyfus, H. (1992). *What Computers Still Can't Do.* Cambridge, Mass.: MIT Press.

Dreyfus, H., & Dreyfus, S. (1988). *Mind over Machine.* New York: Simon & Schuster.

Ekbia, H. (2008). *Artificial Dreams: The Quest for Non-Biological Intelligence.* Cambridge, U.K.: Cambridge University Press.

Fodor, J. (1980). Methodological Solipsism Considered as a Research Strategy in Cognitive Science. *Behavioral and Brain Sciences, 3*, 63-73.

Hempel, C. (1998). Two Basic Types of Scientific Explanation. In M. Curd, & J. Cover, *The Philosophy of Science: The Central Issues* (pp. 685-694). New York: W. W. Norton.

Hofstadter, D. (1985). Waking up from the Boolean dream, or, subcognition as computation. In D. Hofstadter, *Metamagical Themas* (pp. 631-665). New York: Basic Books.

Hofstadter, D. R. (1979). *Godel, Escher, Bach.* New York: Basic Books.

Knuth, D. (1973). Computer science and mathematics. *American Scientist, 61*(6), 707–713.

[*] This is not to say that one couldn't train a computer to recognize "old-fashioned sounding" poems via machine-learning techniques. It's just that the algorithm—say, a neural network—responsible for the program's performance would be pitched at a "sub-personal" level; it would not be an algorithm that a person could easily follow (and certainly not in a reasonable amount of time).

Leder, H., Belke, B., Oeberst, A., & Augustin, D. (2004). A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology, 95*, 489–508.

Reber, R., Schwarz, N., & Winkielman, P. (2004). Processing fluency and aesthetic pleasure: Is beauty in the perceiver's processing experience? . *Personality and Social Psychology Review, 8*(4), 364–382.

Shanker, S. (1998). *Wittgenstein's Remarks on the Foundations of AI.* London: Routledge.

Shanon, B. (2008). *The Representational and the Presentaional.* Charlottesville, Va.: Imprint Academic.

Simon, H. (1995). Artificial intelligence: An empirical science. *Artificial intelligence, 77*, 95-127.

Turing, A. (1950). Computing Machinery and Intelligence. *Mind, 59*(236), 433–60.

Wimsatt, W. (2007). *Re-Engineering Philosophy for Limited Beings: Piecewise Approximations to Reality.* Cambridge, Mass.: Harvard University Press.

Wittgenstein, L. (1967). *Lectures and conversations on Aesthetics, Psychology, and Religious Belief.* London: Basil Blackwelll.

Wright, C., & Bechtel, W. (2007). Mechanisms and psychological explanation. In P. Thagard, *Philosophy of Psychology and Cognitive Science (Volume 4 of the Handbook of the Philosophy of Science)* (pp. 31-79). New York: Elsevier.