6-2022

What's So Artificial and Intelligent about Artificial Intelligence? A Conceptual Framework for AI

Rebekah L. H. Rice
Seattle Pacific University

3

What's So *Artificial* and *Intelligent* about Artificial Intelligence?

A Conceptual Framework for AI

Rebekah L. H. Rice

There is currently a good deal of attention being focused on artificial intelligence, broadly speaking, and deep learning, specifically. The attention is warranted, as these technologies are predicted to affect our collective lives in innumerable ways even beyond their already expansive social reach. There is much to consider regarding the benefits and potential harms of AI. And of course there are the apocalyptic musings about super-intelligent machines running amok, bringing science fiction scenarios uncomfortably close to anticipated reality. But productively engaging in discussions about the ethical and social implications of AI, and about which sorts of futures it is reasonable to anticipate, requires clarity about certain key concepts at play in these conversations. Some of these are conspicuous: *artificial* and *intelligence*, notably. The former suggests a contrast with some other concept. But which, exactly? *Natural*, perhaps? Or *organic*? And *intelligence*, being as it is regularly attributed to human persons, might suggest a fitting analogy with intelligence as it occurs in you and me. But what is intelligence in humans? Does it require a mind or a soul? Is it simply a corollary of electrical and chemical processes in the human brain? Gaining clarity about the range of meanings to which such terms refer—and familiarity with the relevant

debates surrounding the various meanings—will provide the conceptual framework necessary to better articulate the precise ethical and pragmatic questions we think most important to our efforts to intentionally navigate a world with AI.

I. THINKING MACHINES?

When mathematician Alan Turing published "Computing Machinery and Intelligence" in 1950,[1] logical positivism permeated the intellectual terrain.[2] Central to logical positivism is the verifiability criterion of meaning.[3] According to it, the content, or meaning, of any meaningful statement is exhausted by the conditions that must be verified to obtain if the statement is to be considered true. These verification conditions are either empirical (synthetic) or logical (analytic). Those who held the view claimed that if a statement cannot be verified using logic or empirical investigation, then it is unverifiable and therefore meaningless (and not a candidate for truth). Logical positivism ultimately fell out of favor given its implication that many seemingly understandable statements about such topics as metaphysics and religion and political theory turn out to be meaningless. But even more damning, the view's central claim appears to fail to satisfy its own criterion. After all, can the verifiability criterion of meaning itself be verified in the manner required?

What is perhaps most illuminating about this intellectual movement for our purposes, however, is the way in which the logical positivists traded metaphysical questions for epistemological ones. Verifiability is an epistemic criterion. It has to do with what one can know, reasonably believe, or demonstrate. And yet, might there be facts that obtain though no one does, or perhaps can, come to know them? Take the claim that certain mathematical entities—sets, say—are real. Can this be verified empirically? Or can the reality (or not) of sets be established via logical inference? It would seem not. And yet it may nevertheless be that sets belong, as Bertrand Russell contends, to "the world of being."[4] The relevance to Turing is evident when one considers the question with which he begins his seminal paper and, correspondingly, the method he suggests for answering it. The paper sets out to

1. Turing, "Computing Machinery and Intelligence," 433–60.

2. See, e.g., Ayer, *Logical Positivism*; Carnap, "Elimination of Metaphysics." For an application to the mind, see Carnap, "Psychology in Physical Language," in Ayer, *Logical Positivism*.

3. Correspondingly, there is also a verifiability criterion of truth.

4. Russell, *Problems of Philosophy*, 91–100.

answer the question, "Can machines think?"[5] But rather than rehearse proposed definitions or catalog how the terms "machine" and "think" are used, he proposes his now famous imitation game. Here's how he describes it:

> It is played with three people, a man (A), a woman (B), and an interrogator (C). . . . The interrogator stays in a room apart from the other two. The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. He knows them by labels X and Y, and at the end of the game he says either "X is A and Y is B" or "X is B and Y is A." The interrogator is allowed to put questions to A and B. . . . The ideal arrangement is to have a teleprinter communicating between the two rooms.[6]

The interrogator's objective is to ask questions of A and B and then determine, based on their respective answers, with whom he is communicating. It is A's aim to mislead the interrogator and veer him toward an inaccurate identification, whereas B's aim is to help the interrogator. But how, in an imagined game such as this, do we make headway toward answering the question about whether machines can think? Turing clarifies:

> We now ask the question, "What will happen when a machine takes the part of A in this game?" Will the interrogator decide wrongly as often when the game is played like this as he does when the game is played between a man and a woman?[7]

Now, instead of identifying which interlocutor is the man and which the woman, the interrogator must determine when he is receiving responses from a machine and when he is interacting with a human. Could the machine's responses successfully imitate those of a human so that the interrogator believes he is communicating with a human person? If yes, then we will thereby have an affirmative answer to the original question. In other words, a machine that can perform certain kinds of tasks—in this case, answering pointed questions—in a way that is indistinguishable from a human performing such tasks satisfies Turing's criterion for a thinking thing. Notice it is an observer's ability with respect to distinguishing between the two that establishes whether or not the thing is thinking.

Now, that may be well and good and we might think the substituted questions adequate to the task of determining the status of such a machine. But that will depend on what assumptions we are making. Importantly, the

5. Turing, "Computing Machinery and Intelligence," 433.
6. Turing, "Computing Machinery and Intelligence," 433.
7. Turing, "Computing Machinery and Intelligence," 434.

original question is a metaphysical one. It asks whether reality is such that there are, or could be, thinking machines. Or perhaps, more pointedly, it asks about a certain category of items, namely machines, and whether they are the sorts of entities capable of thinking. Now, such a question can be answered in the manner suggested only if thinking is entirely a publicly observable phenomenon. That is, only if thinking is entirely a matter of certain inputs being followed by particular behavioral, or otherwise observable outputs, is the question about what the interrogator is able to observe and conclude on the basis of such observation apt. If the observer compares the behavioral outputs of the machine with those of the human and finds them qualitatively indistinguishable, then just as the human is a thinking thing (*ex hypothesi*), so is the machine. In other words, whether Turing's proposed method is capable of answering his original question depends on what thinking is.

Indeed, the logical positivists who gave thought to the nature of the mental opted for a version of behaviorism fittingly termed "logical behaviorism." According to this view, "all sentences of psychology describe physical occurrences, namely, the physical behavior of humans and other animals."[8] The upshot is that any meaningful psychological statement, that is, a statement purportedly describing a mental phenomenon, can be translated, without loss of content, into a statement solely about behavioral and physical phenomena. Famously, Carl Hempel suggested that "Paul has a toothache" can be translated into a sentence like, "Paul weeps and makes gestures of such and such kinds."[9] The view is that psychological statements ultimately reduce to mere "motions and noises." Now, it's not difficult to appreciate the motivation for reducing psychology to behavior. Behavior, as we've said, is publicly available for observation, whereas internal mental states like pains, beliefs, and desires are not. As such, behavior can serve as an intersubjective verifiability requirement. This facilitates something we should want, namely, the ability for psychological statements to have public, sharable meanings that serve as vehicles of interpersonal communication. What's more, a meaning's being sharable is critical to our ability to analyze it, make generalizations with respect to it across multiple subjects, and potentially treat psychological conditions. It is perhaps no accident that logical and other forms of behaviorism took root during the early part of the twentieth century, rather immediately on the heels of the recent emergence of psychology as a robust field of scientific inquiry and study.

8. Carnap, "Psychology in Physical Language," 107.
9. Hempel, "Logical Analysis of Psychology, 17.

It is helpful to note here that behaviors, which are by their very nature observable and therefore useful because they are verifiable, are importantly distinct from actions. A behavior is whatever people or organisms (or even mechanical systems) *do* that is *publicly observable.* In humans, these can include physiological reactions and responses (e.g., perspiration, salivation, increased pulse rate), and bodily movements (e.g., raising an arm, flipping a light switch, uttering a sentence). In a computing system, they might result in digital outputs of various kinds. Actions, on the other hand, are behaviors typically performed intentionally and for reasons. As such, they are mentally quite robust. To get a sense of the difference between the two, consider Donald Davidson's description of his morning's events:

> This morning I was awakened by the sound of someone practicing the violin. I dozed a bit, then got up, washed, shaved, dressed, and went downstairs, turning off a light in the hall as I passed. I poured myself some coffee, stumbling on the edge of the dining room rug, and spilled my coffee fumbling for the *New York Times*.[10]

Some of the events Davidson describes are actions. Others are mere behaviors—events that simply befall him. Among those belonging to the first category are getting up, washing, shaving, and turning off the light. Those belonging to the second category include being awakened, stumbling on the rug, and spilling the coffee.

Consider the action of turning off the light. From the perspective of an observer, the viewable bodily movement is a flipping of a switch. That movement is followed, presumably, by the room's coming to no longer be illuminated. But what about this behavior makes it a turning-off-of-the-lights? If Davidson had been mistaken about the purpose of the switch, he might have flipped it intending to bring about some entirely different result—perhaps he'd intended to run the garbage disposal. So, at the time that he flips the switch, we, as observers, can't know what action Davidson takes himself to be performing simply by observing his behavior. Or to put matters more precisely, which action Davidson performs depends in part on his reasons for performing it. Suppose I am walking back and forth from one end of the room to the other. What am I doing? Am I pacing? Am I exercising? My legs are moving so as to carry me from one location to another. But that doesn't settle what it is I'm doing—what action I am performing. What action I am performing depends, in part, on *why* I am behaving as I am.

In the case of Davidson, to discover which action he is performing we would need to know what it is he saw in acting—in this case, in flipping the

10. Davidson, *Essays on Actions and Events*, 43.

switch—such that it seemed to him the thing to do. We would require access to his reason or reasons for acting. Reasons and their ilk are mental items. And unless mental items reduce entirely to bodily movements, or some other publicly available phenomenon, it would seem that observation alone cannot help us here. So, when it comes to "full-blooded actions," of the sort that humans (and perhaps members of some other species) regularly perform, it is arguable that a view like behaviorism is likely to fall short.[11] And this is because such actions involve a richer psychological structure than a view like logical behaviorism can account for.

The insistence that all psychological items can be reduced to behaviors marked a significant departure from the way psychological states were historically characterized. Following on a tradition heavily influenced by René Descartes, William James states that "Psychology is the Science of Mental Life, both of its phenomena and of their conditions. The phenomena are such things as we call feelings, desires, cognitions, reasonings, decisions, and the like."[12] Now, if the items of psychology are those James lists, then they are distinctively mental items. Mental items are those that are available to their subjects via introspection, such as the mood I'm now in. Or the thought you're now having. How do you and I come to know these things about our respective inner mental lives? When literal descriptions fail, we turn to metaphor. Introspection, we might say, is the exercise of turning one's gaze inward. It is direct and unmediated.[13] Metaphors aside, one need simply attend to one's own mental states, and *voila*! Introspection allows one to access what is available from the first-person perspective. I cannot introspect and hope to access your mental states, and neither can you so access mine.

Now imagine you wish to study the psychological items that James describes. Because such items are available only from the first-person perspective, it will be difficult indeed to draw conclusions about a broad category of human psychology. Any inquiry I engage in will have a sample size of one (me). That is hardly grounds for a reliable generalization! It is no surprise, therefore, that some have wished to reduce psychological states to behaviors, neurophysiological activity, and the like, being as these are occurrences observable from the third-person perspective. And as such, they lend themselves nicely to serving as the subjects of scientific inquiry. But can the mental be so reduced? Beliefs, desires, pains, perceptual states, and

11. The distinction between low-level activities (e.g., a spider that manipulates its limbs so as to move across the floor) and full-blooded actions is found in Frankfurt, "Problem of Action."

12. James, *Principles of Psychology*, 1.

13. Notice that my access to your mental states is indirect and must be mediated by your testimony or behavior.

so on have long been characterized as items belonging to the mind—or, as Descartes described it, to the part of a person that *thinks*.

II. MINDS AND BRAINS

Whether facts about human psychology ultimately reduce to physical facts—about physical outputs like behaviors, or else those pertaining to our underlying neurophysiology—is a question that has consumed philosophers of mind for a number of decades. Neuroscience tells us there is a high degree of correlation between mental phenomena (beliefs, desires, perceptions, sensations, intentions, and so on) and brain phenomena (neural events, chemical processes, and the like). Indeed, in many cases brain science can tell us which mental states (or events) correlate with which brain states (or events). Couple this with the fact that additional discoveries are being made at a seemingly ceaseless pace, and it is natural to suppose that we can eventually, with time, come to have an exhaustive list of such correlations. In other words, the empirical data makes reasonable the belief that for every mental state *M*, there is a physical correlate, *P*.[14] Of course, this acknowledgment won't by itself reveal precisely how the mind relates to the underlying physiology. But additional facts regarding the apparent causal connections between the two may bolster the case. For example, we know that if we increase the availability of certain chemicals in the brain (e.g., serotonin), it will affect the subject's mood. And we know that damage to certain regions of the brain will result in memory loss, or impairment of speech. So, the connection between the mental and the physical appears to be quite tight—so tight, in fact, that many of the phenomena we once attributed to the mind are now routinely explained by appealing solely to goings-on in the brain. It's not an enormous leap to conclude, on this basis, that the mental just is the physical, or, in any case, that it reduces to the physical, or that it utterly depends, in some other way, on the physical. Indeed, it appears to be a methodological assumption in some disciplines that if one cares to understand the mind, the thing to do is to examine the brain.

Two broad views concerning the relationship between the mind (or the mental) and the brain have emerged. They are substance dualism and physicalism. The first has enjoyed lengthy historical prominence, only to be surpassed in popularity by the second relatively recently. It is to these that I now turn.

14. This chapter admittedly does a good deal of hand-waving with respect to the neural correlates of mental states. For more, see Baker's chapter in this book, "Reinforcement in the Information Revolution."

II.a. Substance Dualism

According to the substance dualist, every human person is composed of two substances: a nonphysical mind (or soul), on the one hand, and a physical body, on the other.[15] The view is Platonic in its origins and received considerable development by Descartes. In Descartes's view, while it's true that I have both a mind and a body, the thing I am—and that which accounts for my continued existence over time—is my mind. As Descartes puts it, "But what then am I? A thing that thinks. What is that? A thing that doubts, understands, affirms, denies, is willing, is unwilling, and also imagines and has sensory perceptions."[16] The mind, then, is the substance in which all of the mental states and activities reside. From a Cartesian perspective, the mind and the body causally interact, as when my feeling hungry causes me to reach for the Cheetos, or when the tissue damage from a tumble off my bike produces a pain sensation in me. This feature of Descartes's view gave rise to a challenge by one of his contemporaries, Princess Elizabeth of Bohemia, in which she asked how it is that "the human soul can determine the movement of the animal spirits in the body so as to perform voluntary acts—being as it is merely a conscious (*pensante*) substance."[17] What Elizabeth seems to be pointing to is the inadequacy of the sort of mechanistic view of causation that (it would have been thought at the time) appears to account fairly well for physical-to-physical causation to account for mental-to-physical causation. And she wonders whether Descartes can offer an alternative given that he countenances mind-body causal interaction. As it turns out, Descartes cannot, and the mind-body problem has come to plague the sort of dualism Descartes defended ever since.

Just what "problem" the mind-body problem exposes for substance dualism remains a topic of dispute among philosophers. In particular, philosophers disagree about whether it reveals a deep incoherence in the very idea of mind-body causal interaction. But there's no question that substance dualism has fallen out of favor in philosophical circles. Interestingly, Elizabeth herself claims that she "could more readily allow that the soul has matter and extension than that an immaterial being has the capacity of moving a body and being affected by it."[18] She would sooner abandon dualism in favor of physicalism than attribute causal efficacy to an immaterial substance.

15. The terms *mind* and *soul* will be used interchangeably throughout to refer to a nonphysical part of persons, if any there be.

16. Descartes, *Meditations on First Philosophy*, 83.

17. Descartes, *Correspondence with Princess Elizabeth*, 53.

18. Descartes, *Correspondence with Princess Elizabeth*, 55.

Replying to the mind-body problem became a central task of Cartesians in the years following Descartes. And while it would be an error to minimize the role the problem played in whittling away at enthusiasm for the view, dualism was more likely eclipsed by physicalist theories over the course of the last century for empirical reasons. It is the demonstrably tight connection between the mental and the physical that rendered a view committed to the independence of my mind and my body (Descartes imagined it possible for me to exist without my body) untenable. And it is the ability to supplant previous appeals to an immaterial soul with explanations in terms of brain functioning that has boosted the plausibility of a physicalist alternative.

II.b. Physicalism

The majority of contemporary philosophers today are physicalists. The same is true of a good number of biologists, physicists, psychologists, and neuroscientists, as well, I understand, as a growing number of theologians.[19] Physicalists believe that human beings, like you and me, are composed entirely (and only) of physical stuff. The implication, of course, is that, contrary to what Descartes and Plato thought, you and I are neither wholly nor partly constituted by an immaterial soul. Instead, I just am my body. Or, perhaps more accurately, I am some part of my body, most probably some part of my brain or central nervous system. Consider a particular human person, Vanessa. If physicalism is correct, then Vanessa is through and through a material, or physical, entity. Now, it won't do to insist that Vanessa is identical to her whole body, since some parts of Vanessa's body could go missing (she could lose an arm in an unfortunate accident, for example) and Vanessa would nevertheless continue to exist. Suppose Vanessa is to undergo a radical transplantation surgery in which a significant number of her critical organs—her heart, lungs, liver, and kidneys, say—are to be switched out for new ones. I suspect Vanessa will be quite nervous about the upcoming surgery, but it's unlikely she will wonder who will exit the operating room upon the surgery's completion. Of course it will be Vanessa because none of those particular organs are essential to her being Vanessa. On the other hand, imagine Vanessa is instead facing a brain transplantation surgery in which

19. N. T. Wright puts his own view this way in "Mind, Spirit, Soul, and Body": "Just as I believe that we are wrong to look for a god-of-the-gaps, hiding somewhere in the unexplored reaches of quantum physics like a rare mammal lurking deep in the unexplored Amazon jungle, so I believe we are wrong to look for a soul-of-the-gaps, hiding in the bits that neuroscience hasn't yet managed to explain."

the surgeon will remove her brain and replace it with someone else's. Now it would seem reasonable for Vanessa to be quite concerned about just who will be wheeled out of the operating room. All of this is simply to suggest that if we are physical things, likely some parts or features will matter more than others. Some physicalists deny this. They claim instead that Vanessa is not identical to her brain (or any part of her body), but to a living organism.[20] Living organisms routinely lose bits (by, say, shedding skin cells) and acquire new ones, assimilating them into the complex system in a way that preserves the organism's existence, provided the replacement occurs gradually. But as these considerations regard a person's persistence over time, we needn't settle them here. What is central to the current discussion is that a physicalist, of any variety, claims that human persons are identical to an entity that is physical through and through.

Physicalism comes in a couple of forms. Reductive physicalists maintain that a person's mental life (i.e., her beliefs, desires, intentions, emotional states, and so on) is wholly reducible to neural events and chemical processes in her brain. The upshot is that there's nothing distinctive or special about the mental. The mental just is the physical. One concern about reductive physicalism is that it appears to be incompatible with the thesis that human beings sometimes act freely. After all, reductive physicalism implies that everything I do is caused by neural events in me. Suppose I raise my hand because I wish to hail a taxi. According to reductive physicalism, my hand's rising is not caused by my desire (to hail a cab), but rather by some neural event which sends a signal (ultimately) to my limb. But neural events, like all physical events, are governed by physical laws. And which physical laws there are, and whether they hold, are not matters that are up to me. So, it seems that my actions are not up to me.[21]

Nonreductive physicalists maintain, like their reductionistic counterparts, that there are only physical substances and that I am one such. But they deny that mental states can be reduced to brain states. That's because a belief, for example, cannot be wholly described, without loss of meaning, in purely physical terms. Among other things, beliefs have intentional content. They are "about" something. Take, for example, my belief that Seattle is in the state of Washington. My belief is about Seattle. But the corresponding neural event (whatever that may be) is not about anything at all; it is merely a biological state. And the same is true of desires, intentions, and perceptual states. The mental is anomalous—truly unique and irreducible. Notably,

20. See, e.g., van Inwagen, *Material Beings*; Olson, *Human Animal*; Merricks, "Resurrection of the Body."

21. An astute reader will note that free will is precluded on this picture only if free will is incompatible with determinism.

these mental items inhere in, or are states of, the physical item (the body).[22] In this way, nonreductive physicalists deny the substance dualist's insistence that mental items must inhere in a mental substance (the mind).

To avoid reduction, nonreductive physicalists characterize the relationship between the mental and the physical in a way that upholds a dependence of the former on the latter, but which avoids identity. The options are rather abundant, as several such relations have been proposed.[23] One way to unpack this dependence is in terms of supervenience. There are several versions of supervenience on offer, but for our purposes it will do to express the idea in terms of the well-known maxim, "no mental difference without a physical difference." In other words, if *x* and *y* are in every way alike physically, then they are in every way alike mentally. Importantly, the sort of dependence envisaged here is asymmetric (the mental is dependent on the physical, but the physical is not similarly dependent on the mental).

An attractive feature of nonreductive physicalism is that it preserves our understanding of action. For nonreductive physicalists, the mental is causally efficacious and able to bring about the bodily movements that count as our actions. Whereas the reductive physicalist will reduce the causal efficacy of the mental to physical causal relations at the subvenient base, nonreductive physicalists claim that the mental is itself causal. My running really is caused by my seeming to see a lion, and your eating the chocolate is indeed brought about by your desire to do so. But can the nonreductionist help herself to mental causation? A well-rehearsed argument, known as the "Exclusion Argument," suggests not.

Jaegwon Kim articulates the argument this way. Begin with a metaphysical doctrine likely to garner sympathy from any physicalist: the causal closure of the physical domain (or "closure principle," for short).

> Closure Principle: If a physical event has a cause at time *t*, it has a sufficient physical cause at *t*.[24]

In searching for causes of physical events, we never need venture beyond the realm of the physical.[25] As Kim puts it, "the physical domain is

22. A closely related view which characterizes these irreducible mental items as properties (rather than states or events), but nevertheless denies the existence of a mental substance, is property dualism.

23. Proposed relations include (but are not limited to) constitution, emergence, realization, and supervenience.

24. Kim, *Philosophy of Mind*, 214.

25. Indeed, Kim goes on to say that "if closure fails, theoretical physics would be in principle incompletable" and that "it seems clear that research programs in physics, and the rest of the physical sciences, presuppose something like the closure principle."

causally, and hence explanatorily, self-sufficient and self-contained."[26] Suppose, as our above story about action implies, that a mental event, *m*, causes a physical event, *p*. It follows from this and the closure principle that there is also a physical event, call it *p**, occurring at the same time as *m* that is a cause of *p.* To preserve nonreduction and hence the causal efficacy of the mental event *m*, we will need to posit that *m* is not identical to *p**. But now we have two purported causes of *p*: *m* and *p**. Unless this is a genuine case of overdetermination, it would seem that *p* leaves little (i.e., no) work for *m* to do. Indeed, as Kim puts it, "No event has two or more distinct sufficient causes, all occurring at the same time, unless it is a genuine case of overdetermination."[27]

Genuine cases of overdetermination occur when two independent causal chains converge at a single effect as when a house fire is caused by a short circuit and a lightning strike simultaneously, or when two bullets hit a person at the same time, either of which would have been sufficient to kill him. We can allow for some causal overdetermination (surely such occurrences *can* happen), but one would expect them to be rare. However, if every case of mental-to-physical causation involves (at least) two sufficient causes, then every case in which I act will be a case of overdetermination. Now multiply this by all actions performed by persons at any time in history and the overdetermination will be very widespread indeed![28]

Embracing this result renders mental states epiphenomenal, or causally inert, and undercuts the familiar account of action with which we began. To vindicate *m* as a genuine cause of *p*, *m* should be able to bring about *p* without there being a synchronous *p**. But in any version of physicalism, every mental event has a physical causal partner (or correlate) that would have brought about the effect, even if *m* had not.

Kim takes the lesson of the exclusion argument to be that, insofar as we wish to preserve mental causation, we must reduce *m* to *p**. Our understanding of agency can be maintained with the proviso that it is not my belief, desire, or intention, but rather the respective state's physical substrate (or realizer, if you prefer), which causes the bodily movement that constitutes my action. Actions therefore involve physical causal sequences through and through. The upshot is to deny that the mental is something

Philosophy of Mind, 215.

26. Kim, *Philosophy of Mind*, 214.

27. Kim, *Philosophy of Mind*, 216.

28. Some have argued that the variety of overdetermination involved in cases of mental causation are innocuous. See Bennett, "Why the Exclusion Problem Seems Intractable"; "Exclusion Again"; Sider, "What's so Bad about Overdetermination?"

over and above the physical. Or, put differently, if one is a physicalist, then one should be a reductive physicalist.

Much ink has been spilled over the last two decades in efforts to reply to the exclusion argument and preserve mental causation. These attempts take us too far afield form our present purpose to warrant explication here. Suffice it to say that there is likely no view about the nature of the mind that comes free of cost. As they say, there are no free lunches. And yet one of these, broadly construed, is likely (roughly) correct. It is not my aim to argue for one over the others. But which view one leans toward will determine how one thinks about the possibility of thinking machines—or thinking things of any kind.

III. MINDS AND MACHINES

In this final section, I want to draw out some implications of what has thus far been said for certain questions we might have with regard to AI. Naturally, this is not an exhaustive set of questions. And I don't profess to provide answers. Instead, I take the following (admittedly brief) discussion to be instructive for thinking about how to go about answering some of the vexing questions that arise with respect to artificial intelligence.

Let's begin by returning to Turing's original question: Can machines think? In short, I suppose it depends on what one means by "think." Since Descartes, thinking has been understood as the defining feature of the mental. Now, if mental items are wholly reducible to neuronal occurrences, then the matter becomes largely an empirical one. The same is true if mentality reduces to behavior. Both nonreductive and dualistic theories deny the reducibility of mental items and accordingly will insist on conditions that are not available from the third-person perspective but are instead phenomenal, or else that depend on the presence of states with certain sorts of content, as in the case of beliefs, desires, and the like.

Now, I concede that I've taken a rather narrow path to understanding the intelligence component of artificial intelligence. Of course, a great many things can be meant by "intelligence." I have focused on thinking, much as Turing did. But other concepts like processing, learning, understanding, and inferring may also be relevant. I don't have space to attend to each of these here. This is likely no great loss since the phrase "artificial intelligence" has largely been deemed imperfect from the outset and demonstrably overreaches the sorts of actual technological developments thought to fall under its umbrella. Most of these have specific and more fitting names—for example, machine learning, symbolic systems, big data, supervised learning,

and neural networks. That said, I am less concerned here with the precise technologies than I am with the ambitious questions that have occupied many thinkers when they've sought to imagine what might be possible.

And this brings us to the matter of artificiality. This too might have several meanings. I deem a comparison with grand concepts like *natural* to prove fruitless since it is notoriously difficult to define what is and isn't natural in a way that renders intuitive results (by including the right things and excluding all the others). But here's a way of getting at the sort of question I take many to have when they think about thinking machines. Let's first take a detour back to mental phenomena. Mental states are often thought to be multiply realizable. Consider pain, for example. Pain in humans supervenes on (or perhaps it is identical to) certain processes that occur in the human brain. But octopus pain—a phenomenon I have every reason to believe is quite real—is realized by drastically different physical occurrences (owing to its different physiology). Moreover, if one thinks it coherent to imagine an extraterrestrial being with an altogether unique underlying physical structure to anything found on earth nevertheless experiencing pain, then pain is a concept that is definable independently of the structure that realizes it. Perhaps other mental states—even thinking—are like this. But, then, that's precisely what is at issue in the various theories about the metaphysics of mind we've considered.

But now we might wonder whether the examples we've so far given bear something essential in common. Is it relevant that extraterrestrials and octopuses and humans are composed of material stuff that cannot be produced in a lab or a factory? It's not obvious why it should be. Just as we are wont to ask upon which biological structures mentality might supervene, we may also ask whether mentality can supervene on synthetic structures. To answer in the affirmative is to carve out a metaphysical possibility. It is not to commit ourselves to a view about the physical possibility of such an occurrence nor, certainly, is it to stake a claim with respect to how close technological advances are to actualizing this possibility. Interestingly, it isn't even to articulate an account of what an underlying structure must be like to give rise to mentality. It is to do something quite different. It is to start with the mental. It is to ask what mentality *is*. What does it require? And what, precisely, is the relationship between it and the physical? It likely won't come as a surprise that these are deep philosophical questions about the metaphysics of mind. And while I believe we can make progress with respect to them, they will undoubtedly remain perennial questions.

In the interim, here is a strategy we might employ: At the very least, let's begin by making our assumptions about the mind explicit. If, in your view, thinking requires an immaterial soul, then only entities to which

such an item can be attached (or from which such an item can emerge) will make the cut. If, on the other hand, you believe that mentality depends on physicality, then how so? Does the former reduce to latter, or does it merely supervene on it? Once these commitments are front and center, we can have a robust conversation regarding what sorts of entities are capable of satisfying the conditions of mentality. But if these presuppositions are allowed to lurk in the background and out of plain view, we will almost certainly talk past one another.

The downsides to this are quite real. Could we create a machine such that powering it off would be immoral? The moral requirements with respect to the proper treatment of a being depend rather significantly on what sorts of experiences it can have. My children's first pet was a betta fish. It once (quite unintentionally) went unfed for a staggering nine days. How egregious a moral failure this was depends, in part, on the extent to which the lack of food caused the fish to suffer. Even if it had died (it didn't), the loss of life wouldn't necessarily, all by itself, have constituted a moral failure. I've killed many a house plant by not watering it. Few will contend that killing a plant is immoral. Now, I'm fairly confident that betta fish have rather limited sets of experiences. A dog would have fared much worse and undoubtedly would have suffered greatly, certainly enough to make the omission morally unconscionable. (This is precisely why a fish makes a better pet for busy households with young children and preoccupied parents than does a dog.) It is also why we (properly) feel less remorse when squishing a spider than we do harming a cat. Of course, we could be wrong about what sorts of mental states a thing enjoys and thereby be incorrect in our moral assessments. Regardless, the metaphysical facts ground the truth of the relevant moral claims. As such, the metaphysics ought not be ignored—and we can begin by being transparent about what metaphysical presuppositions we bring to the relevant discussions.

I'll conclude with a final matter of importance. When the topic of the possibility of thinking machines arises, many begin to wonder if there is something distinctive, and important, about humans such that our unique value is not undermined by the presence of machinery capable of completing many of our tasks and endowed with the ability to think and have experiences. Christian theists, in particular, are apt to worry about the doctrine of *imago Dei*, according to which human persons (or so it's often understood) bear the image of the divine. This has sometimes led theists to view an immaterial soul as a particularly attractive feature by which to mark a human person. This is too quick. First, it's not at all clear that the

imago Dei doctrine is best understood ontologically.[29] Second, if an immaterial soul is the substance that underlies all mental states, then anything exhibiting mentality will *ex hypothesi* have a soul. But now the motivation for the original appeal to the soul—that it secures the distinctness of human persons—has vanished.

It's also noteworthy that within Christianity there is robust debate about whether substance dualism is a view properly understood to be suggested by Scripture and creedal doctrine, or is instead a product of philosophical (notably, Platonic) influence.[30] Christians, and indeed theists generally, might find dualism less objectionable than their nontheistic counterparts given their view that God created, and continually interacts with, the physical universe. In this way, they already allow for instances of nonphysical-to-physical causation. It is perhaps largely for this reason that Christian physicalists have not tended to list the mind-body problem among their primary motivations for denying dualism and adopting physicalism. Even so, a growing number of Christian philosophers and theologians reject substance dualism.[31] And so it would not be accurate to regard an appeal to the soul as the distinguishing mark of a human person as an essential Christian commitment.

Theistic views aside, a being's value needn't depend on its status as ontologically or phenomenologically distinct from other existent beings or entities. And while it may be true for any entity that it either has mental states or lacks them (dolphins have them, flower pots lack them), beings that have mental states can differ from one another considerably (the inner mental life of a dolphin is likely quite different from that of a bat). Were we to discover thinking extraterrestrials, it wouldn't thereby follow that human mental phenomena would in any way be diminished. Nor would this be true were it to turn out to be possible for certain synthetic structures to manifest thought. Such considerations are worth bearing in mind as we consider the question about wherein mentality can reside.

29. For an alternative conception of the doctrine, see De Cruz and Smedt, "*Imago Dei* as a Work in Progress."

30. For a range of views on this topic, see Murphy, *Bodies and Souls*; Wright, "Mind, Spirit, Soul, and Body"; Rickabaugh, "Dismantling Bodily Resurrections Objections to Mind-Body Dualism"; Lugioyo, "Whose Interpretation? Which Anthropology?" and Cooper, "OK, But Whose Misunderstanding," in Crisp et al., *Neuroscience and the Soul*.

31. Notable examples include Baker, *Persons and Bodies*; Corcoran, *Rethinking Human Nature*; Merricks, "Resurrection of the Body"; Murphy, *Bodies and Souls*; O'Connor, *Persons and Causes*; van Inwagen, *Material Beings*.

BIBLIOGRAPHY

Ayer, A. J., ed. *Logical Positivism.* New York: Free, 1959.

Baker, Lynne Rudder. *Persons and Bodies.* Cambridge: Cambridge University Press, 2000.

Bennett, Karen. "Exclusion Again." In *Being Reduced: New Essays on Reduction, Explanation, and Causation*, edited by Jakob Hohwy and Jesper Kallestrup, 280–307. New York: Oxford University Press, 2008.

———. "Why the Exclusion Problem Seems Intractable, and How, Just Maybe, to Tract It," *Nous* 37 (2003) 471–97.

Carnap, Rudolph. "The Elimination of Metaphysics through the Logical Analysis of Language." *Erkenntnis* (1932) 60–81.

Corcoran, Kevin. *Rethinking Human Nature: A Christian Materialist Alternative to the Soul.* Grand Rapids: Baker, 2006.

Crisp, Thomas M., et al., eds. *Neuroscience and the Soul: The Human Person in Philosophy, Science, and Theology.* Grand Rapids: Eerdmans, 2016.

Davidson, Donald. *Essays on Actions and Events.* Oxford: Oxford University Press, 2001.

De Cruz, Helen, and Jovan D. Smedt. "The *Imago Dei* as a Work in Progress: A Perspective from Paleoanthropology." *Zygon* 49 (2014) 135–56.

Descartes, René. "Correspondence with Princess Elizabeth." In *Modern Philosophy*, edited by Forrest E. Baird, 53–56. Englewood Cliffs, NJ: Prentice Hall, 2008.

Descartes, René. *Meditations on First Philosophy.* In *Selected Philosophical Writings*, 73–122. Cambridge: Cambridge University Press, 1988.

Frankfurt, Harry. "The Problem of Action." *American Philosophical Quarterly* 15 (1978) 157–62.

Hempel, Carl. "The Logical Analysis of Psychology." In *Readings in Philosophy of Psychology*, edited by Ned Block, 1–14. Cambridge, MA: Harvard University Press, 1980.

James, William. *The Principles of Psychology.* New York: Henry Holt, 1890.

Kim, Jaegwon. *Philosophy of Mind*. Boulder, CO: Westview, 2011.

Merricks, Trenton. "The Resurrection of the Body." In *The Oxford Handbook of Philosophical Theology*, edited by Thomas P. Flint and Michael C. Rea, 476–90. Oxford: Oxford University Press, 2009.

Murphy, Nancey. *Bodies and Souls, or Spirited Bodies?* Cambridge: Cambridge University Press, 2006.

O'Connor, Timothy. *Persons and Causes: The Metaphysics of Free Will.* Oxford: Oxford University Press, 2000.

Olson, Eric. *The Human Animal: Personal Identity Without Psychology*. New York: Oxford University Press, 1997.

Rickabaugh, Brandon. "Dismantling Bodily Resurrection Objections to Mind-Body Dualism." In *Christian Physicalism? Philosophical Theological Criticisms*, edited by Keith R. Loftin and Joshua R. Farris, 295–317. Lanham, MD: Lexington, 2018.

Russell, Bertrand. *The Problems of Philosophy.* Oxford: Clarendon, 1912.

Sider, Ted. "What's so Bad about Overdetermination?" *Philosophy and Phenomenological Research* 67 (2003) 719–26.

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

van Inwagen, Peter. *Material Beings.* Ithaca, NY: Cornell University Press, 1990.

Wright, N. T. "Mind, Spirit, Soul, and Body: All for One and One for All Reflections on Paul's Anthropology in his Complex Contexts." Presented at the Society of Christian Philosophers Regional Meeting, Fordham University, 2011. http://www.ntwrightpage.com/Wright_SCP_MindSpiritSoulBody.htm