JOURNAL OF SOCIAL COMPUTING
ISSN 2688-5255 03/07 pp275−325
Volume 4, Number 4, December 2023
DOI: 10.23919/JSC.2023.0029

AIdeal: Sentience and Ideology

Daniel Estrada*

Abstract: This paper addresses a set of ideological tensions involving the classification of agential kinds, which I see as the methodological and conceptual core of the sentience discourse. Specifically, I consider ideals involved in the classification of biological and artifactual kinds, and ideals related to agency, identity, and value. These ideals frame the background against which sentience in Artificial Intelligence (AI) is theorized and debated, a framework I call the AIdeal. To make this framework explicit, I review the historical discourse on sentience as it appears in ancient, early modern, and the 20th century philosophy, paying special attention to how these ideals are projected onto artificial agents. I argue that tensions among these ideals create conditions where artificial sentience is both necessary and impossible, resulting in a crisis of ideology. Moving past this crisis does not require a satisfying resolution among competing ideals, but instead requires a shift in focus to the material conditions and actual practices in which these ideals operate. Following Charles Mills, I sketch a nonideal approach to AI and artificial sentience that seeks to loosen the grip of ideology on the discourse. Specifically, I propose a notion of participation that deflates the sentience discourse in AI and shifts focus to the material conditions in which sociotechnical networks operate.

Key words: sentience; agency; artifacts; artificial intelligence; ideology; nonideal theory; natural kinds; participation

1 Deflating Sentience: A Cynical Approach

This paper offers a cynical and deflationary perspective on sentience and its application to Artificial Intelligence (AI) and robotics. My perspective is cynical because in this historical moment I am pessimistic about our collective ability to resolve our disagreements and confusions regarding concepts like “sentience”, especially as they pertain to AI. In other words, this project was begun knowing it would fail. My perspective is deflationary because I do not think the irresolvable nature of the discourse has much of anything to do with the “mystery” of sentience, consciousness, or the mind generally. On the contrary, while there remain many things we do not understand about minds and brains, I believe that the broad outlines of a generally correct theory have been widely understood for decades. For instance, the sensorimotor pathways that lead from the detection of an itch to the initiation of a scratch have been understood for many years, including its basis in genetics and its prevalence among animals[1, 2]. While such examples of course do not exhaust the capacities of the mind, they do provide a suggestive framework for understanding its operation, especially regarding sensory capacities and behavioral dispositions broadly shared among living creatures, like the experience of being itchy. Granting some general epistemic humility and in full recognition that we still have much to learn, it is not inaccurate or overzealous to say that many of the philosophical and metaphysical puzzles which mystified ancient and early modern thinkers about the relationships between sensation and action are now textbook parts of well-established evolutionary biology, neuroscience, and cognitive psychology. Nevertheless, there exists little

• Daniel Estrada is with the Department of Humanities and Social Sciences, New Jersey Institute of Technology, Newark, NJ 07102, USA. E-mail: estrada@njit.edu.
∗ To whom correspondence should be addressed.

Manuscript received: 2023-07-28; revised: 2023-11-24; accepted: 2023-12-21

consensus in the literature on precisely what sentience refers to or how it figures into scientific, metaphysical, ethical, or political practices. From my perspective, this absence of consensus speaks less to the complexity of sensory cognition in animals and more to the complicated ideological and political work that the term “sentience” is expected to perform in this historical moment. These complications are exacerbated in the context of artificial agents built by large corporations seeking to maximize profits and neuter regulatory oversight. In the field of AI, the term “sentience” is used not just carelessly but maliciously, for purposes of disinformation and obfuscation. The exploitation of these persistent confusions and ambiguities does not reflect lively philosophical debate so much as it reflects a crisis of ideology. In this paper, I address a collection of ideological tensions involving the classification of agential kinds, which I see as the methodological and conceptual core of the sentience discourse. Specifically, I will consider ideals involved in the classification of biological and artifactual kinds, and ideals related to agency, identity, and value. Together, these ideals frame the ideological background on which sentience in AI is currently theorized and debated, a framework which I call the AIdeal. To make this framework explicit, I will begin by reviewing the historical discourse on sentience, paying special attention to how these ideals are projected onto artificial agents. Our goal in this exercise is not to resolve the issue of artificial sentience directly; on the contrary, we will soon see that this issue has been resolved again and again in the history of ideas. Instead, our goal is to appreciate how the classification of artificial agents and the possibility for artificial sentience has historically been fixed by broader philosophical and ideological commitments. Indeed, the status of artifacts often figures explicitly in attempts to articulate those broader commitments, typically as a contrast with “genuine” human agency. Reviewing these historical commitments and their categorical implications will provide some critical distance on the ideals at stake in contemporary iterations of these well-worn debates, where many of these ideals continue to thrive. Our historical review will proceed as follows. In Section 2, I offer some preliminary remarks on sentience as marking a distinction between plants and animals as biological kinds. I will consider both the ancient origins of this idea and its plausibility in contemporary taxonomy. In Section 3, I consider Aristotle’s account of sentience as the faculty of the soul characteristic of animals. Aristotle’s account is rendered in a metaphysics of teleological essentialism in which the relationships between artificial and biological kinds are relatively clear. Moreover, the essentialist intuitions motivating Aristotle’s account are often implicit in the discourse around artificial sentience today. For these reasons, it will be instructive to render this account in detail. This section ends by considering the possibility for artificial sentience on an Aristotelian account. The ideals of early modern philosophy were partly characterized by a critical rejection of Aristotelian teleology, and partly by an embrace of the mechanical explanations championed by the developing scientific revolution. The fundamental confusions and apparently irresolvable tensions in contemporary debates are largely an artifact of the historical transition from ancient to early modern accounts of minds and machines. In Section 4, I explore these early modern ideals through the work of Descartes, Hume, and Bentham. Descartes’ dualist account of mechanical bodies and immaterial minds radically shifts the categorical boundaries not only between humans and animals but also between agents and artifacts. Hume and Bentham further develop the ethical and political dimensions of these distinctions. In Section 5, I consider the possibility for artificial sentience on early modern accounts. I will argue that the transition from ancient to modern to contemporary understanding of the mind has resulted in a discursive situation where it is simultaneously necessary and impossible for an artifact to be sentient. I argue that this contradiction is at the core of our ideological crisis today. One upshot of this historical review will be to disabuse ourselves of a certain naive teleological presumption in the artificial sentience discourse. One might have thought that while ancient thinkers could only dream of artificial minds in their myths and stories, as time passed and technology improved, we have continued to make incremental progress towards this once distant goal. From this naive perspective, one might interpret the recent flurry of scholarship calling for more rigorous terminology in artificial sentience as evidence that we are closing in on the goal, where such precision would be necessary. Our historical review should make clear that this narrative of slow but steady progress towards artificial sentience is a self-aggrandizing myth. Artificial sentience has never been a stable goal that can be pursued in some objective sense, the way we might attempt to land on the Moon or cure cancer. On the contrary, the discourse on artificial sentience has the structure of an endless hallway or a carrot on a stick, where the goals are persistently just over the horizon or just around the corner, depriving us of any clear measure of progress no matter how much ground we seem to cover. If we could somehow fix the goalposts in this debate to the ideals of our predecessors, it might reveal dozens of Moon-landing scale events of technical achievement in the last fifty years that would easily convince Aristotle, Descartes, or even Turing that an artifact had achieved sentience. Scholars today would yawn at the same examples. The point is not to ridicule ourselves or the intellectual tradition we have inherited, the point is simply to acknowledge that our attitudes towards artificial sentience have changed so dramatically over time that it is impossible to see this history as aimed at any coherent goal that has not already been achieved many times over. Put simply, the prospect of artificial sentience depends less on what artifacts do, and much more on what we believe about them. That is the AIdeal. Having laid out this history, in Section 6, I critique the AI discourse as an exercise in what Charles Mills[3] called "'ideal theory' as ideology". Specifically, the AI discourse posits both machines and humans as abstract agents with idealized capacities that can be compared and evaluated in absolute terms through abstract measures like "intelligence". The contemporary discourse on artificial sentience reinforces this deep seated commitment to ideal theory, a tendency found even among AI's sharpest critics. Following Mills, I argue that the way through this debate is not by attempting to resolve a consensus among competing ideals, but instead by shifting focus to the material conditions and actual practices in which these ideals operate. In this spirit, I sketch a nonideal approach to artificial sentience. I argue that much of the scholarly hand-wringing in the last year over mentalistic language applied to generative chatbots is driven by a concern for ideals and ideology to the exclusion of the actual. This critique cuts both ways. It cuts against the shameless industry hype parroted by wide-eyed tech journalists. It also cuts against the reactionary humanist's desperate attempts to draw sharp lines between humans and machines. We are thus left in need of a perspective that treats artificial sentience as neither impossible nor inevitable, but instead as a persistently problematic and janky part of our actual world, a routine fixture of the everyday lives of actual people. In Section 7, I conclude the paper by sketching a notion of participation that parallels and complements the ideological structure of agency as it figures in the sentience discourse. Participation differs from agency as such in that it centers group rather than individual activity. Using the formalism of membership in fuzzy sets, I argue that participatory frameworks allow for consideration of the diverse and interdependent networks of activity from which agency emerges, and thus operates as a necessary component of any ethical theorizing of agency. A participatory approach to the AIdeal allows for a kind of pragmatic flexibility in our recognition of artificial agency and sentience, and thus can accommodate the myriad ways these artifacts show up in our actual lives. Throughout the discussion, I use the term "discourse" in a nontechnical sense to mean something like "conversation". By "popular discourse", I am usually referring to popular entertainment (books, movies, and TV) and the various communities and media they generate. I will sometimes identify "tech journalism" as playing a specific role in the popular discourse, for instance. When I use the term "historical discourse", I am typically referring to historical science and philosophy, and I will try to be clear about which historical periods I am referring to. By "AI discourse", I typically mean academic research in AI and robotics, and especially in AI ethics, the discipline in this field I know best, but also including Science and Technology Studies (STS) and Human Computer Interaction (HCI) research into the social impact of these technologies. As this all suggests, the field of AI is notoriously interdisciplinary, with fuzzy boundaries not just with other academic research but also with other industries, including popular media. Thus, I will often name the "contemporary discourse" as an inclusive umbrella term for both academic and popular engagement with AI, robotics, and artificial sentience since, say, the start of the century, and especially since the recent AI boom began around 2012. I will also use the term "agent" or "agency" in a general, non-technical sense to refer to anything that *does something*. Agents are basically dynamical objects. Whereas an object is (potentially) inert and lifeless, an agent performs some operation, and therefore interacts with the world in some way. A metronome is an agent; its operation is to generate a ticking sound on a regular interval. Agents can be mechanical devices, but they can also be software agents that exist merely as snippets of simple code. This notion of agency does not assume mental faculties beyond what explains its operation. Thus, I will use the term "artificial agent" to refer to artifacts (things built by people) like chatbots and robots, and the term "artificial sentience" to discuss the general question of sentience in artifacts.

2 Nutritive Plant and Sensitive Animal

In this section, I look at sentience as an ancient tradition for marking the distinction between plants and animals. I consider the relevance of this distinction in contemporary biology and taxonomy. I also consider whether sentience distinguishes between natural kinds, discussing the implications of this question for plant and machine sentience. Finally, I reflect on categorical logic and the politics of essentialism as foundational structures of the AIdeal.

2.1 Nutrition and sensation

Sentience is the capacity for sensation. Sentient creatures are sensitive to changes in their world, and might deploy a variety of strategies to detect changes in light intensity, color, temperature, pressure, motion, electrical charge, and chemical composition, among many other sensory modalities. The motivating question in philosophical and scientific debates over sentience is how to understand these varieties of sensitivity and the mechanisms that underlie them. A broad understanding of sensitivity might include any responsiveness to environmental change, which leads some thinkers to conclude that sentience is a ubiquitous characteristic of living creatures, from humans and other animals to chemotaxis in single-celled microbes. Some go further by noticing that a variety of relatively simple artifacts are also sensitive and responsive to environmental changes, such as a liquid thermometer or the sensor-triggered movement of an automatic door. Even so-called "inanimate" material in the natural world can be reactive to surrounding conditions, as when iron rusts in the presence of oxygen or lithium reacts in water. If "sentience" is sensitivity to environmental change understood in the broadest possible sense, perhaps it includes the whole of the universe and every piece of it, as every molecule and particle, every quark and electron and neutrino, is reactive and responsive to the things around it—albeit in different ways and to different degrees. While this totalizing view represents one extreme in the spectrum of views on sentience, it should not be confused with panpsychism, as it does not require some rich internal conscious states for each particle. Instead, this is the extreme pan-interactionism of plain old-fashioned physics, that familiar and thoroughly externalist science that recognizes the contingency of things on other things, processes on other processes.

Most thinkers who find the word "sentience" useful tend to use it in a more restrictive way. The heart of the sentience debate concerns which restrictions to take on board and where to draw the important boundaries. There are compelling reasons for drawing lines in many different places, often aligned with certain evolutionary and biological distinctions among living creatures. It is increasingly popular to associate sentience with the activity of a nervous system, a particular biological structure characteristic of animals and which underlies most of their sensory and motor functions. However, it would be a mistake to conclude too quickly that sentience names a specific anatomical or physiological distinction. The historical roots of the discourse on sentience are found in ancient theories on biological life, specifically regarding the distinctions between plants and animals. The philosophical discussion traces to Aristotle's writings on biology, ethics, and his theory of the soul in *De Anima*, itself a variation of the tripartite soul in Plato's *Republic*; but of course these and earlier texts did not invent the distinction between the "nutritive" (or "vegetative") plants and "sensitive" (or "sentient") animals. Instead, they try to explain a distinction that everyone took to be obvious and could see for themselves. Plants are alive; they need water, sunlight, and seasonal conditions to grow and reproduce; they are clearly responsive to changes in their world. Plants can also die, both as part of their life cycle but also through disease, consumption, and mistreatment. Together, these "nutritive" faculties of growth, reproduction, and death were understood to be common among all plants, and indeed shared among all living creatures. Animals likewise live, grow, reproduce, and die, developing the particular needs and interests of their kind along the way. However, unlike plants, animals are directly responsive to immediate environmental change. A flower does not flinch before you step on it, but even a simple fly will avoid your swat. To say that animals are "sentient" has historically named this apparent distinction in sensitivity and responsiveness between plants and animals; Aristotle says "it is sensation primarily which characterizes the animal"[4]. The crucial difference, again taken by Aristotle and others as common sense, was that animals seek to avoid injury, whereas plants do not seek to avoid anything, even potentially fatal injury. We know in our own case that injuries can be painful, and that pain motivates our avoidance of further injury. It is a simple inference to conclude that plants do not sense injury or feel pain the way we and other animals do. Since plants will stay put while you cut them down, and might even grow back in the same place to be cut down again the following season, they must not experience these interventions the way animals do. So, while plants and animals are both alive and responsive to their environments, animals have a unique form of sensitivity, one that is associated with direct perception of environmental change, and especially with sensations of pain. At least, this is a sketch of the traditional narrative and justifications, inherited from the ancient world, for describing the living creatures as categorically divided between the nutritive plants and sensitive animals.

Before trekking further into the sprawling issues around this narrative, we should pause to take stock of its prominent features. The concept of sentience arises in the traditional narrative to describe the distinctions between plants and animals that were apparent to the ancient world. As such, the term inherits many of the confusions and misunderstandings of the biology of that period. Modern plant biology reveals plants to be engaged in myriad forms of sensitive interaction with their environments, including on short time scales, and in some cases using electrical signaling similar to nervous systems in animals. While some interpret these findings as evidence for sentience in plants[5], the more immediate and uncontroversial conclusion is that the apparent distinctions in capacities between plants and animals are not nearly as clear as we once thought. For this reason, it is tempting to treat "sentience" as an anachronistic holdover of those ancient traditions, and to look for new words and framings grounded in modern science to characterize our improved understanding of these issues. Evolutionary biology has to some degree moved away from distinguishing creatures by capacities and morphology, and today attempts to ground taxonomic distinctions in genetic relatedness. We know from genetic analysis that plants and animals are distinct clades with a most recent common ancestor over 1.6 billion years ago[6]. Whatever their confusions, the ancient world was clearly correct to notice deep and systematic differences between plants and animals.

Nevertheless, the historical divergence of these clades does not determine the characteristics they might have in common. Good tricks are regularly rediscovered by natural selection, as with the independent emergence of flight in insects, birds, and bats. A biologist in the ancient world might have wondered what insects, birds, and bats have in common that allows for the shared capacity of flight; indeed, until Linneas' modern taxonomy was introduced in 1735 biologists often classified bats as a kind of bird, since warm-blooded flight was a defining characteristic of birds as a category, and bats are warm-blooded flying creatures[7]. We know today that while these creatures are all animals and share an evolutionary lineage, it is not a shared genetic heritage or physiological trait that allows them to take to the skies. The capacity for flight is not explained by an intrinsic or essential characteristic that all and only flying creatures share. Instead, what these creatures share are body plans that successfully manage the ratio between lift and drag, the physical parameters that constrain all things from becoming airborne, including airplanes, umbrellas, and ourselves. Flight is a way of dealing with the constraints and dynamics of the physical world—air resistance, buoyancy, and turbulent flow—that to some degree all living creatures must contend with. Natural selection has repeatedly found effective strategies for managing these constraints that allow some creatures to remain in the air for extended periods; indeed, there are many distinct strategies for doing so, each with their own risks and advantages. Plants, too, have found diverse strategies for developing seeds or even entire body plans that can remain in the air for extended periods. Biological flight and other cases of convergent evolution are worth dwelling on because they provide clear cases for thinking about the awkward ways our categorical intuitions map (or fail to map) onto the complex biological landscape. Flying snakes and flying squirrels both flatten their bodies to glide through the air; the similarity of this behavior has little to do with common features between the snakes and squirrels, and far more to do with the aerodynamic complications of jumping quickly between trees. The lesson of such examples is that what at first seems like an intrinsic similarity between creatures (perhaps made apparent by a naming convention) might in fact be the result of extrinsic similarities of the task environment that these creatures are confronting across evolutionary time.

2.2 Natural kind

In the philosophy of science, we can sum up these considerations by saying that flying creatures do not form a natural kind. A basic challenge in the sentience discourse is a lack of clarity over how natural kinds and categorical logics operate, so we will take the time to work through the instructive example of flying, and we will return to it repeatedly in this section. Undoubtedly many distinct creatures fly, and have done so long before the evolution of our earliest primate ancestors. But flying creatures do not form a "natural kind" in the sense of a set distinguished by some intrinsic or essential property that is exclusively characteristic of its members. Instead, the set of flying creatures is a haphazard mish-mash of kinds because flying is ultimately the result of extrinsic pressures that many different creatures are subject to, and therefore does not reflect some intrinsic or essential property shared by those creatures. The set of flying creatures does not "carve at nature's joints", as the saying goes. As far as I know, it is not controversial even among committed believers in natural kinds to say that flying creatures are not an example of a natural kind. Other examples of sets that are not natural kinds are the set of yellow things, the set of things taller than five meters, and the set of things in Colorado. In each case, while the members of these sets include natural things in the actual world, they do not pick out natural kinds.

The traditional narrative takes for granted that sentience marks a boundary between natural kinds. For the moment we might set aside the mistakes and confusions of ancient biology, and ask in a modern context if it makes sense to continue treating sentience as distinguishing a natural kind. There are many views on natural kinds and a rich literature discussing their purchase in science and policy[8−10]. I do not intend to settle these issues or even substantially weigh in on them in this paper, but some introductory distinctions will be helpful in what follows. There are two major lines of thought on natural kinds: realism and constructivism[11]. Realists believe that natural kinds identify real distinctions in nature. Put simply, nature has joints, or brute distinctions and categories, and perhaps some of our natural kind terms track those joints reliably. Constructivists believe that scientific terms and distinctions are constructed through scientific practice, and therefore to some extent exist as social constructions. Both views come in strong and weak varieties. Strong realists believe that the world is ontologically arranged into kinds, and the scientific study of nature requires cataloging and investigating its kinds. Weak realists are less committed to drawing strong metaphysical conclusions from scientific practice, but they share the realist view that science to some extent studies the distinctions it finds in nature, distinctions that are "really there" in a mind-independent sense. Strong constructivists hold that no distinctions between kinds are given in nature, and that all apparent kinds are artifacts of human social activity projected onto the world. Weak constructivists hold that perhaps there are some mind-independent distinctions in the world, but our attempts to investigate and theorize these distinctions are inherently filtered through human social practice, and must be understood in the context of those practices. My impression is that the majority consensus of modern scientific opinion regarding natural kinds is somewhere in the union of weak realism and weak constructivism, which for our purposes can be treated as a single coherent view. For instance, to say that biological flight is not a natural kind is compatible with both weak realist and weak constructivist accounts of natural kinds. For that matter, it is also compatible with views that reject the notion of natural kinds altogether.

Does sentience name an intrinsic distinction between natural kinds? In the traditional narrative, this question is equivalent to asking whether sentience marks the distinction between plants and animals. The question has two parts: Are plants and animals natural kinds? And does sentience name the distinction between these kinds? Given a broadly inclusive sense of natural kinds, the case for animals and plants as natural kinds seems straightforward; if anything in biology is a natural kind in the intuitive sense, plants as distinct from animals is as good a case as any. However, this case can be made independent of any appeal to sentience. Animals as a phylogenetic clade have many characteristics that we might take as markers of an intrinsic distinction in kind from plants. Animal cells lack a cell wall and do not typically use photosynthesis to drive metabolism. Animals have nervous systems, which plants characteristically lack. Plants and animals as groups have many genetic markers of distinct evolutionary pathways going back for more than a billion years. If one is motivated by the language of natural kinds in either a weak realist or weak constructivist sense, these characteristics are more than sufficient for distinguishing plants from animals as natural kinds. They are also uncontroversial parts of high school science education. While sentience figured in ancient accounts of the distinction between plants and animals, we can today draw a more precise version of what is basically the same intuitive distinction without appealing to sentience at all. This makes it possible to accept that plants and animals are both natural kinds, but to reject the idea that sentience names an intrinsic distinction between them. This is effectively the position of arguments that plants are sentient. These arguments do not argue that plants are animals, or that no distinctions exist between these kinds. Instead, they claim that sentience is not limited to the animals, and therefore cannot be counted as an intrinsic feature of animals in a sense that is exclusive of plants. Rather, sentience is something closer in kind to flying: a set of diverse strategies for coping with real-world challenges found among many living creatures, not a reflection of some intrinsic characteristic of any specific category of living creature.

It is reasonable at this point to treat the philosophical discussion as resolved. The ancients were basically correct that plants and animals are distinct kinds, but they were perhaps a bit hasty to treat sentience as marking the intrinsic distinction between these kinds. If this is correct, there is no deep challenge to recognizing many different kinds of things, including sociotechnical kinds, as potentially capable of sentience, just as we recognize planes, kites, and umbrellas as sociotechnical kinds capable of flight. Once we agree that sentience does not distinguish a natural kind, we might continue to haggle over the details of definitions in particular contexts for particular creatures, but we have accepted that the issue is fundamentally a matter of choice. There is not much point in arguing over whether flying snakes “really” fly, because there is no single thing that flying “really” is, and therefore no ultimate criteria for determining a correct answer. Likewise, if sentience does not distinguish a natural kind, then there is no point in arguing over whether artificial kinds are “really” sentient, because there is no single characteristic that sentience “really is”. There is a well-known quote in AI from Dijkstra[12] expressing this perspective: “The question of whether a computer can think is no more interesting than the question of whether a submarine can swim.” Dijkstra is not saying that computers cannot think; he is saying that thinking (like swimming) does not distinguish between natural kinds, so it is futile to expect some definite resolution to the question one way or the other. Remember, denying sentience as a natural kind distinction does not require denying natural kinds as such. It also does not require denying that any living creature is sentient, just as denying flying as a natural kind does not imply that no creature flies. This is not a statement of illusionism, or strong realism/constructivism about natural kinds, or any other controversial metaphysical thesis. This conclusion is driven most directly by the simple fact that in modern biology a distinction between plants and animals can be grounded in ways that do not appeal to sentience as a distinguishing characteristic.

In any case, we are now in a position to examine a tension between two claims that might appear at first glance inconsistent but, given the above considerations, can both easily be true at the same time:

- Sentience in animals is explained by a nervous system.
- Plants are sentient but do not have a nervous system.

We do not require a complicated theory of sensory cognition or consciousness to reconcile these claims. They are easily compatible as long as we do not treat sentient creatures as forming a natural kind. If sentience does not distinguish between natural kinds, plants and animals can both be sentient in different ways, just as insects and birds fly in different ways. We can consistently believe that despite their differences, plants and animals are both “really” sentient, in the same way that, despite their differences, insects and birds both “really” fly. Flight in insects and in birds both deserve serious scientific investigation, and these fields are independent of each other in the sense that neither study takes priority or can make direct claims against the other. The legitimacy of insect flight does not require any comparison with bird flight; structural dissimilarities between insects and birds do not count against either creature’s capacity to fly. It would be absurd to use research on the neural basis of insect flight to mount an argument that birds cannot actually fly, or vice versa; both creatures obviously fly, even if they go about it in different ways. A tight analogy holds for plant and animal sentience. If sentience does not distinguish between natural kinds, the analogy also extends directly and without qualification to artificial sentience. These claims only become controversial if we are committed to thinking about sentient creatures as a natural kind of the sort that can exclude other kinds. We will discuss the motivation for such views in the following sections.

Nevertheless, the contemporary literature in science, ethics, and policy is overwhelmingly committed to treating sentience as marking an important boundary between natural kinds, to debating membership claims for specific creatures, and to investing this boundary with significant moral and political weight. This commitment is apparent even within biological and ecological ethics entirely unconcerned with AI and other sociotechnical kinds. Defending sentient creatures as a natural kind is taken to require a difference in essence or capacities that is distinct from and potentially more restrictive than the basic cladistic distinction between plants and animals. As mentioned earlier, the activity of a nervous system is widely treated as the most likely biological basis for sentience. Nearly all animals have nervous systems except for creatures like sponges, animals in the phylum Porifera that have nerve cells as larva but which consume these cells when they have found a good place to settle as adults. Adult sponges appear to live a largely sedentary and plant-like lifestyle. It is not hard to see sponges as a relic of an evolutionary history where plants and animals were less distinct creatures; conceptually, sponges act as a boundary case between categories that highlights their distinctions. Platypuses likewise serve a conceptual role in understanding the distinctions and evolutionary relatedness between mammals like us, and egg-laying, bill-having creatures like birds and reptiles. Conceptualizing these creatures in this way is not exactly accurate or fair; sponges and platypuses have been evolving alongside other animals the whole time, and the creatures we see today reflect this evolutionary history as much as we reflect our own. Nevertheless, sponges provide a convenient example of creatures that are phylogenetically animals but (at least as adults) are supposedly not sentient *because* they lack a nervous system, thereby justifying sentience as a *distinct* distinction in kind between plants/animals, a distinction that “really exists” in nature, giving legitimacy to its use in ethics and politics. Identifying sentience with a nervous system, which we might as well call the Identity Thesis, allows us to keep the category distinctions in *almost* the same place as the ancient theories, minimally disrupting our long-held intuitions while providing a convincing biological basis for them. It lets us have the ancient cake of sentience and eat it in a modern biological setting. Endorsing the Identity Thesis has become popular among a community of bioethicists identifying with the term Metazoan[13], which is an archaic term for animals from a 19th century classification system distinguishing the proper animals from the Protozoa, another archaic term for single-celled eukaryotic lifeforms that lack cell walls. It is worth explicit reflection that this classification system emphasizes that animals as more closely related to single-celled protists than to plants. In the context of increased politicization of sentience as a basis of ethical consideration, this framing is an overtly political act. To put things bluntly, the term “Metazoa” as it is used in bioethics today is anti-plant propaganda.

This political posturing against plants might be surprising if the stakes of this debate only involved a distinction between natural kinds. After all, it is no problem to be an identity theorist who treats sentience as a characteristic of nervous systems, while at the same time rejecting sentient creatures as a natural kind. This is no more problematic than treating flight as a characteristic of feathered wings, while at the same time rejecting flying creatures as a natural kind. I have argued in this section that rejecting sentience as a natural kind comes at little practical, inferential, or epistemic cost, especially for everyday practices of scientific research and science communication. Rejecting sentience as a natural kind is compatible with a broad range of mainstream views on the metaphysical status of natural kinds. What we get in return for stepping over these minor theoretical hurdles is that many of the apparent incongruities and philosophical puzzles between plant, animal, and artificial sentience dissolve into essentially bureaucratic details and we can all get on with our lives. Of course, what has been left out of this discussion so far is any engagement with sentience as an experiential quality of intrinsic ethical concern. Addressing these issues in moral theory and philosophy of action will occupy Sections 4 and 5.

2.3 Logic and politics of classification

Before moving on, we should note a few other issues regarding form and politics in the classification of kinds. It will be useful to consider Aristotle's influential framework a bit more explicitly, as nothing we have said so far has required it. Aristotle makes several foundational contributions to a systematic thinking about life and the mind. One is his hylomorphic causal theory, the famous "four causes" that lay out a structural relationship between matter, form, and function, which we will return to in the next section. Another is a categorical system of logic, which he took to be a fundamental part of scientific inquiry. For Aristotle, biological kinds can be organized into categories divided by some distinguishing characteristics. These categories can be arranged into a logical structure, part of what today we call set theory, that allows inferences with universal or existentially quantified statements establishing their relationships, such as:

- All animals are sentient.
- No plants are sentient.
- Therefore, no plants are animals.

One can easily give a rigorous proof that this argument is formally valid; the truth of the premises guarantees the truth of the conclusion. Aristotle used the term *episteme* to describe a kind of science that involves the application of categorical arguments like these to deduce features of the natural world. In this approach, we discover facts about nature by reflecting on the relationships between the categories we find in nature and the logical relationships between them. In other words, categorical distinctions are features of the natural world that serve as a basis for reasoning about its structure. The distinctions are given by nature, and we can use categorical logic to study their properties and relationships, thereby generating new knowledge about the world. Nevertheless, the use of categorical logic requires care! Consider the following argument, consisting of the same claims as above but slightly rearranged:

- All animals are sentient.
- No plants are animals.
- Therefore, no plants are sentient.

This argument is invalid; the premises fail to support the conclusion, which is to say that both premises can be true while the conclusion is false. Such examples suggest that natural science requires more than just attention to the distinctions we find in nature; it also requires paying careful attention to the formal inferences we draw from these distinctions. We must think about *how* we are thinking, not just what we are thinking about. These are early struggles in a long philosophical tradition emphasizing the formal structure of thought over its practical import.

Of course, too much emphasis on formal considerations carries its own share of pitfalls. Russell famously criticizes Aristotle's reasoning from claimed categorical distinctions in size and strength between men and women to the false conclusion that women have fewer teeth than men. Russell[14] quips, "although he was twice married, it never occurred to him to verify this statement by examining his wives' mouths". Aristotle's biology is undoubtedly wrong in many details and misogynistic in its perspective[15], but Russell's anecdote misrepresents Aristotle as insensitive to the inferential ambiguities of categorical logic. In reality, Aristotle's biology repeatedly highlights examples where distinctions cut across categories. For instance, Aristotle distinguished the animals on land into two-legged and four-legged creatures, warm- and cold-blooded creatures, and between creatures that lay eggs and creatures that give birth to live young. He elaborates that some two legged creatures lay eggs (birds) while some give live birth (humans); some four-legged creatures lay eggs (reptiles) and some give live birth (cows and horses), and so on[16]. The implication of these examples is that there is no uniform way to group the categories, and no all-encompassing hierarchy for organizing these distinctions that can be derived from first principles. Theophrastus, a student of Aristotle's who wrote the *Historia Plantarum*, the plant-focused companion to his teacher's *Historia Animalium*, remarks at the beginning of this text that while plants have parts that are distinguished by function (flowers, leaves, roots, etc), the way these parts develop and contribute to reproduction can be very different from the organization of animals. He continues: "And in general, as we have said, we must not assume that in all respects there is complete correspondence between plants and animals"[17]. It is clear that biologists in Aristotle's day recognized that the careless appeal to categorical inferences could lead us astray. To his credit, Aristotle did not make the "mistake" of classifying bats as birds. Instead, he explicitly recognized bats as an example of a creature that "dualizes" between categories, that is, they "belong in one way to one group and in another way to another"[7]. Aristotle describes bats as flying creatures that give birth to live young and have other characteristics of typical four-legged creatures. Aristotle points to ostriches as another category-straddling case, which lays eggs and has feathers but does not fly; these are the ancient analogs of sponges and platypuses in modern taxonomy. This nuanced appreciation for categorical relationships is ironic given the appearance of cosmic inevitability in the *scala naturae*, the Great Chain of Being, often viewed as originating with Aristotle's biology and which influenced thinking about biological kinds for the next millenia[16]. Aristotle did see categories as divided between superior and inferior kinds, with humanity superior to all other animals, a uniquely perfect form of living creature. Aristotle also believed that women as a category were less rational than (and so inferior to) men[18, 19], and that some people were "natural slaves"[20]. Perhaps if he had given as much care to his writings on women and the oppressed as he did to categorical logic, Aristotle's intentions regarding natural hierarchies would not have come across history the way they have.

Contemporary science no longer expects the natural world to conform to our naive intuitions about categorical structure and classification. It is not only plausible but quite likely that the word "sentience" names a distinction that the ancients took to be an obvious fact about our world, but that on further reflection does not exist in any substantive way beyond these appearances. Sentience was a metaphysical best guess our predecessors made thousands of years ago about the nature of plants and animals, one that has so deeply ingrained in our culture, language, and practices that we struggle to think outside its framing. Ancient biologists deserve more credit than they typically get for their innovative thinking about the natural world; this is especially true for the thousands of nameless biologists who over countless generations developed and improved the common sense thinking which Aristotle and his contemporaries could treat as data to be explained. What the ancients failed to appreciate is that some of the deepest distinctions we find in nature are nevertheless an accident of evolutionary and cosmological circumstance, and on their own provide little guidance in understanding the world. If sentience is simply a relic of ancient biology, debating sentience in machines today would have the absurd quality of asking whether motor oil warrants expanding the traditional four humors to five.

And yet the absurdity of this discourse is not benign. Arguments for drawing boundaries on sentience as distinguishing between natural kinds reinforce an anachronistic method of sorting nature into kinds, with uncomfortable implications of biological essentialism. Biological essentialism is the idea that living creatures have an essential and immutable ("God-given") nature which make them essentially distinct from other kinds. Not only is biological essentialism incompatible with evolutionary theory and with modern approaches to taxonomy and classification, it often serves as an ideological justification for race and gender essentialism, and thus has an acute political valence in contemporary discourse. These political overtones ring more loudly when this classification debate is taken outside the realm of evolutionary biology, as is the case with the popular discourse over sentience in AI and robotics. Regardless of where one draws the boundaries on sentience, contributing to debates in popular media over how to distinguish sentient from non-sentient creatures to some extent provides scientific legitimacy for ignorant or mean-spirited questions about what distinguishes men from women, or trans people from cis people, or black people from white people from asian people. The idea that the business of science is to draw sharp lines around natural kinds, especially when those kinds overlap with sociopolitical or sociotechnical kinds, is both methodologically dated and politically regressive, and it has no place in contemporary scientific discourse. To the extent that the sentience debates in AI are an exercise in reasoning about sociopolitical categories as natural kinds, perhaps it should be discouraged and avoided altogether.

If the concept of sentience derives from a traditional distinction between plants and animals, it is also worth considering the extent to which this traditional distinction serves as a measure or template for evaluating novel claims of sentience, especially in the case of robots and AI. When evidence for sentience in plants is presented, that evidence is weighed not only on its merits, or by some standard checklist for animal sentience. It is also evaluated by whether the evidence overcomes the convenience of existing classification schemes, and the inconveniences of reclassification. Linnaeus' reclassification of bats from the birds to mammals comes at a conceptual and epistemic cost; for instance, it constrains the sort of inferences I can easily make. If bats are birds, I can confidently point to any warm-blooded (that is, non-insect) flying creature and *know* it is a bird, whether or not I can distinguish birds from bats. If bats are *mammals*, and mammals are not birds, then all such inferences lose confidence, and I have to be more careful with my claims about the flying creatures around me. To overstate a perhaps obvious point, people knew long before Aristotle that bats are hairy and give birth to live young. These are two features that today we consider characteristics of mammals that distinguish them from birds. But it was not the case historically that we discovered these facts and then realized our classification error; these facts were well known the whole time. What changed was the weight we give these features in our classification schemes. Linnaeus did not simply reclassify bats as if moving a label from one shelf to another. He introduced novel ways for thinking about how to group the most salient features of living creatures, and indeed conceived of entirely new categories of creature, such as "the mammals", that gave the bat's apparent category-straddling features a "natural" explanation[21]. Of course, these classification issues pervade science and are not limited to evolutionary biology. Pluto's contested status as a planet showcases the same basic tensions between practical and formal virtues in the organizing schema of science.

These considerations raise questions about exactly what we are doing by inquiring into the sentience of machines. Such inquiries assume that the classification schemes concerning sentience are fundamentally legitimate or require minor adjustment to account for specific cases that have potentially been misclassified. But we might just as well challenge the classification schemes themselves, or the project of classification as such. What is the goal of a debate over artificial sentience? Is it to discover and understand some facts about the world? To elucidate some normative consideration or structure? Is it to protect a category of creature through explicit representation in the discourse? Or is it to protect a set of intuitions, inferences, and cultural practices that have been around in some form for thousands of years? These goals are all tangled together, but it is not hard to imagine realistic cases where they come apart, and where we might have to prioritize some over others. It is not the goal of this paper to resolve these issues. I have argued that we should reject sentience as distinguishing between natural kinds, and I have pointed to some reasons to reject it altogether, but I do not think these arguments are decisive and I do not expect them to persuade others.

However, by unpacking these issues we have highlighted one of the central pillars of the sentience debate, which is the assumption that creatures can be classified into kinds and categories that reflect their essential nature. This basic commitment to essentializing categories is a consistent holdover from ancient theories of biology and remains a prominent feature of contemporary sentience discourse. In the next section, we will look at sensation, thought, and movement in order to catalog other ideals at play.

3 Sensation, Thought, and Movement in Aristotle

In this section, I sketch Aristotle's theory of the soul, the faculties of nutrition, sensation, thought, and movement, and the distinction between action and passion. I then consider whether an artificial agent could be sentient on Aristotle's account. I will repeat this exercise in Sections 4 and 5 by sketching an early modern agent, drawing from Descartes, Hume, and Bentham, and considering the possibility of sentient artificial agents on an early modern view. These accounts are not exhaustive, of course, and I do not intend to defend or reject either historical view. Our goal in reconstructing these perspectives is to better understand sentience as a faculty of the mind, the ideals at play in its articulation, and to appreciate how certain complexities and confusions in the contemporary discourse on sentience have their origins in conversations that began thousands of years ago.

As I am writing for a general audience that might not be steeped in the philosophical canon and might not see the immediate relevance of the archaic technical distinctions we will be drawing, some motivating words are in order. Reviewing the debates of historical figures is not simply an academic exercise in classifying philosophical positions or honoring dead white men. While Aristotle, Descartes, and other towering figures in the history of philosophy were real human beings who lived and breathed as we do, their contributions to the history of ideas requires treatment at least somewhat independent of the biographical details of a single individual. The work that bears their names inspired countless other real human beings to produce other works, to invent entire domains of knowledge, to organize their communities, to live worthwhile lives. When philosophers name historical figures, those names are sometimes used as shorthand for the forces of culture, language, and material history made possible by their works, if only indirectly. It is of course unfair to credit these forces to the actions of any individual person, and using names as shorthand for ideas is ultimately a bad habit of writing that raises the barrier for entry into the discourse. The practice is only excusable to the extent that so many of those who contributed to our shared history did so in explicit recognition of the influence these thinkers had on their work. Highlighting Descartes' implicit rejection of Aristotle in the *Meditations* is not merely an amusing quirk in the published diary of a long-dead European. These texts are like seismographs, recording tectonic shifts in the landscape of ideas, shifts that occurred not because these texts were written but because they were taken up and critiqued and engaged by others and woven into the landscape of collective thought. Through his writings and letters, Descartes, the historical person, charts out a small portion of that landscape, highlighting the notable things he finds there, planting distinctive flags that will help later arrivals orient themselves. As time goes on, the word "Cartesian" comes to refer to the territory he charted and the flags he planted, and how they relate to other flags, perhaps charted by figures long after he had gone. The name refers not just to a person but to coordinate directions on a crowdsourced map that generations of people helped to build. Reviewing historical debates is a way of reconstructing these maps for ourselves.

3.1 Sentience as a faculty of the soul

"... the faculties of the mind hunt in packs."

—A. O. Rorty[22]

Sensations are the sensory components of experience. Aristotle's *De Anima* ("On the soul") popularized the idea that there are five senses: touch, sight, hearing, taste, and smell, each tasked with detecting specific features of the world. These sensory capacities together characterize part of the soul's faculty for sensation and perception, or *sentience*, nestled alongside the faculties of nutrition, thought, and movement in Aristotle's organizing schema[20]. The "soul" (*psyche*) is the substance or essence of a living organism; for Aristotle, "having a soul" is synonymous with "being alive". Aristotle was interested in "movement" in a broad sense that includes both locomotion (moving around the world) but also the changes associated with growth, development, reproduction, and decay that are characteristic of all life. These broad patterns of activity are shared by plants and animals, and categorically distinguish them from the "inanimate" material of the natural world. For Aristotle, the inanimate world was composed of the fundamental elements, earth, water, air, fire, themselves emerging from more fundamental dichotomies between hot and cold, wet and dry. Aristotle believed that living creatures are also material beings composed of dynamical arrangements of the physical elements. However, he believed that life is driven or animated by an organizing principle (*energeia*) to seek the goals or ends (*telos*) characteristic of its kind. This organizing principle should not be confused with the "ghost in the machine" we find in early modern philosophy; it is not some distinct immaterial substance that moves matter from within. Instead, the soul is the principle that explains the organization of the matter itself, in terms of the kind of creature it is, and the characteristic features of that kind. For instance, a child might ask "Why do ducks lay eggs?" The Aristotelian response still has an intuitive pull: ducks lay eggs because they are birds, and it is the nature of birds to reproduce by laying eggs. Laying eggs is characteristic of birds as a kind, and since ducks are birds, they too lay eggs. Knowing the animal's kind gives insight into their characteristic activity (*ergon*), which gives context for explaining their behavior and body plans, and how these characteristics contribute to the creature's life and goals.

The soul's faculties or capacities (*dunamis*) are its active powers to move, change, and engage the world in pursuit of a good life. The faculties of the soul are like the organs of an organism, each playing a distinct role in the operation of the whole. Nutrition is the basic faculty of growth, reproduction, and decay shared by all life forms, and which exhaust the faculties of plants. Plants are alive and so they have interests and needs, and they develop characteristics that aim to satisfy their needs and pursue their interests, with the ultimate aim of living a good life (*eudaimonia*). Thus, a plant's roots grow down into the earth and their leaves grow up into the air, because doing so is good for the plant. Animals, including humans, have needs, interests, and preferences in the same way plants do, but we express these interests through the distinct characteristics and faculties of our particular kinds. Sentience is the faculty of perception by which animals sense the world around us and discern its various features. The sensory modalities (such as sight or hearing) are detectors for specific features of the world (such as light or sound). Successful detections produce sensations, which are stirrings of the soul indicating the sensed qualities. Sensations are not free-floating experiences; they are affective, moving the organism into action. Together, nutrition, sensation, and movement exhaust the faculties of the non-human animals.

Thought (*nous*), sometimes translated as reason or understanding, is also a faculty of the soul, but one that is characteristically human. We are *essentially* rational animals; thinking is the characteristic faculty of our kind. The "capacity for reason" is not meant to suggest a strict adherence to formal logic. Although Aristotle did see a rational life as involving the development of good habits and routines, his conception of "reason" did not have the robotic or mechanical connotations it has today. To be "rational" for Aristotle is simply to be sensitive to reasons, to make choices on the basis of deliberation and reflection in pursuit of the good life. Reason does not involve perceiving objects in the world, it involves "seeing" the practical implications of conceptual or logical relationships and factoring them into one's actions. A sensitive animal is moved by their sensations; for instance, a hungry dog will search for food. Aristotle recognizes that some degree of intelligence (memory and planning) is required for any sentient animal to accomplish their goals. However, humans are uniquely rational creatures in that we can be moved by relations of ideas that might themselves have no sensory qualities at all. A hungry dog will eat the food set before it. A hungry person might not eat the food in front of them because they are trying to cut back on sugar, or because they have a dinner engagement later that evening, or because they have ethical reservations with the way the food was prepared, or for countless other reasons. Some of these reasons might be poorly informed, superstitious, or otherwise faulty, but they are all "reasons" of the sort that only humans can be sensitive to. Aristotle's belief that we are "rational animals" does not preclude our tendency to act irrationally.

However, Aristotle admits that neither sensation nor thinking are sufficient to motivate action: "thought by itself moves nothing"[23]. So Aristotle also describes the faculty of movement, the capacity of an organism to move and to be moved by the operations of the soul. Aristotle means "movement" in both a spatiotemporal sense (animals physically move around their world), but also in the affective sense related to the way we might say we are "moved" by a piece of art; movement results from a stirring of the passions. The faculty of movement manifests in animals as desire. Through this faculty we find some sensations or thoughts pleasurable and others painful; desire imbues sensations and thoughts with an attractive or repulsive quality that is ultimately responsible for the complex behavior of all animals. Pain has an experiential component of course, but it is not simply an idle perceptual state, as if receiving a signal from the world through a tin can. On the contrary, pain demands action; pain is a condition to be dealt with. If I put my hand in a fire, the pain will bring me to pull it out quickly; the sensation compels action, even involuntarily. Desire accompanies the sensations of animals in this way, giving the operations of perception and thought an affective quality whose function is to help creatures perform their proper functions and achieve their proper ends. Aristotle identifies three kinds of desires: those pertaining to nutritive and reproductive functions, which he calls *appetite* and which he identifies with sensations of hunger and sexual desire; the moods and emotions like anger or sadness which he calls *spirit*; and desires pertaining to the operations of thought, which he calls *wish* and identifies with the drive to realize the results of deliberation in action, as when one executes a successful plan.

One important background commitment of Aristotle's picture is a principled distinction between actual (*energeia*) and potential (*dunamis*). We can not hope to do justice to these difficult ideas in passing, but since these notions are fundamental to Aristotle's account of the mind, even a rudimentary gesturing in this direction will help with the analysis that follows. We can think of the *actual* as an idealized, fully-realized, perfected state of a thing, which Aristotle associates with formal cause; and *potential* as the inherent power something has to realize that perfect state, which Aristotle associates with material cause. The classic example is an acorn's development into an oak tree. The acorn is alive; its essence carries the form of a fully-realized oak tree which organizes and motivates its material body as it develops from a seed. The acorn's task in life is to achieve that idealized final state in the actual world, to become the best oak tree it can be. The acorn's soul, its organizing principle, fixes the ends it pursues and the faculties it has for pursuing them. So the acorn grows roots and leaves and branches in an effort to become a fully-realized oak tree, its idealized final form (*entelechy*). Those patterns of growth are faculties, *potentials*, tentative developments in material arrangements that contribute to the acorn's active heed of the demands of its soul. These demands necessarily become more complicated in the soul of a sensitive animal. Aristotle's theory of perception, which takes up most of the discussion in *De Anima*, describes how sensory organs detect the properties of objects through some material medium (e.g., vision and hearing operate through air) in such a way that an interaction of material properties allows the transfer or communication of formal properties. Aristotle illustrates the theory with the example of an impression in wax, borrowed from Plato's *Theatetus*. The material interaction of pressing a seal into wax results in a transfer of the form of the seal (its shape) onto the material of the wax, a form which persists after the interaction has completed. In vision, a material interaction between an object and my eye leaves a *formal* impression on the sense organ that becomes available for thought and action. Vision is a capacity, a potential for detection of visible properties in objects, and that capacity is realized whenever I detect that property in some object. This fundamental operation, the change from a state of potential to a fully actualized final state, is for Aristotle the ultimate explanation of all movement in the cosmos.

This metaphysical change from potential to actual is not only central to Aristotle's account of perception, it underlies his theory of agency, and the distinction between action (*praxis*) and passion (*pathe*). Aristotle explicitly associates this distinction with a linguistic difference between active and passive voice: the difference between *stabbing* and *being stabbed*. An action is an active process, initiated by some causal agent exercising its faculties in pursuit of its ends. Each action has a causal counterpart in the thing being acted on; my *action* of chopping an onion has a counterpart in the onion's *passion* of being chopped. Actions are *voluntary* when the soul, the organizing principle, is the formal cause of change, driving the movements of the body. But living creatures are not purely active, they are also passively subject to the processes around them. The *passions*, sometimes translated as emotions, are the forces and processes an agent might be affected by without any voluntary action on their part. So while my feeling of hunger or pain is generated by my body, these are not actions but passions: they are involuntary and I have little control over their appearance. However, this distinction is subtle; Aristotle suggests that the distinction between action and passion is a matter of perspective, giving as an example that a road from Athens to Thebes that is simultaneously a road from Thebes to Athens. Thus, my hunger is a passion that is met with my action of searching for food. So, while passions are involuntary, I am still responsible as a rational agent for how my passions influence and guide my action, and for cultivating a character where my passions do not get out of control.

This superficial sketch of Aristotle's account of the soul is hardly sufficient for any careful work, but for our purposes it will have to do. We can sum things up by defining an Aristotelian agent as having the following characteristics: they are alive, they are self-motivated to pursue category-specific ends, and they are equipped with the faculties of nutrition, perception, thought, and movement, each playing a distinct functional role in the agent's operation. As living agents, they are driven to act in the world; they can also be passively subject to the actions of others, to the passions of their bodies and minds, and to processes in the world beyond their control, all of which might impact their pursuit of a good life. While plants are alive and have a nutritive soul, they lack sensory organs and so cannot perceive the world by detecting its formal properties. Animals are proper agents because they are moved into action in virtue of their desires. To say that animals are sentient refers to both perceptual and affective dimensions of this agency; traditionally the term "sentience" assumes the possession of nutritive faculties and desires/emotions, but does not assume the rational faculties.

Aristotle's theory purports to provide a comprehensive account of the soul in all its biological manifestations: it explains the organization of living creatures and why they work the way they do. It also provides a measure for evaluating how well a creature is doing, relative to the characteristics of their kind. Members of a kind are not cookie cutter copies but admit of variations such that one individual might do well compared to another of the same kind. Moreover, the theory accounts for how an individual might play an active or passive role in different processes, which enables an ethical analysis of an agent's responsibility in that process. We are responsible for voluntary actions that result from rational deliberation, the archetypical case of agency. However, Aristotle makes clear that passions are involved even in deliberative processes and play a constructive role in voluntary action. It is in the context of this account of agency that Aristotle develops his virtue ethics. Virtues are practices and character traits that when cultivated help people realize their ends and achieve a good life. Both desire and practical intelligence play a role in the process of cultivating these virtues. Aristotle emphasizes that virtue is not simply a matter of making a good choice at the right time; virtue is not merely the result of momentary rational insight or good luck. A virtuous person makes good choices as a matter of habit; virtue requires becoming the sort of person for whom good actions and choices come naturally.

Some big picture aspects of this theory are worth highlighting for an AI context. One obvious feature is that Aristotle's metaphysics applies to both material objects and living creatures, and indeed to things that straddle these domains, such as crafted artifacts like an axe or a statue. In fact, Aristotle builds his metaphysical framework and his theory of the soul through direct consideration of artifacts as motivating examples. Artifacts are presented in Aristotle's work as clearly demonstrating the relationships between a thing, its material constitution, its formal structure, its process of development and construction, and its ultimate purpose or function. This same framework applies to artifacts, living creatures, and inanimate objects alike. Artifacts are built to perform their characteristic functions; living creatures also have essential functions, not because they are soulless tools but because they are alive and therefore have purposes that they are intrinsically motivated to pursue. Artifacts are not intrinsically motivated to pursue anything because they are not alive, but they do have "essential natures", and therefore some intrinsic standard by which to judge their performance. For instance, there is an essence to something's being an ax, something that makes it an ax, and it can exemplify those characteristics better or worse. A dull, poorly constructed ax is still an ax, but it is a poor representative of its kind. In this way, the theory recognizes rich analogies between living creatures and artifacts, and develops a unified system of explanation and value that incorporates each domain without collapsing the distinctions between them. Another striking feature of Aristotle's view is that, while living creatures exhibit a unique kind of activity, essential motion drives everything in the cosmos, including inanimate objects. Everything has a purpose, a proper activity that it can approximate more or less well in particular cases. Everything will move of its own internal power to realize this purpose; even the earth itself has a proper place beneath our feet. While

Aristotle does recognize a deep distinction between living creatures and material objects, his view does not locate all action and motion in souls, nor does he treat material objects as inert, dead, or without value. It is a great irony that we struggle to account for artificial mechanical agents today, in the age of mechanical sciences. It is somewhat humbling to appreciate that an ancient theory might be better equipped for theorizing artificial agency than we are.

3.2 Is artificial sentience possible on Aristotle's account?

Whether you sympathize with the view or not, the Aristotelian soul that I have sketched poorly above is a legitimate treasure in the history of ideas, an intricately constructed and marvelously decorated masterpiece of insight into the nature of action, the challenge of living well, and the organized biological processes that make it all happen. Many of the distinctions and nuances of Aristotle's theory of action are obliterated in the modern scientific era; understanding the dynamics of the contemporary discourse requires some appreciation of what was lost along the way. Although many things in Aristotle are archaic and inconsistent with contemporary science, his approach also shares important characteristics with functional or mechanical approaches that fit surprisingly well in the contemporary scientific discourse. It will be a useful exercise to think through some issues in artificial sentience from an Aristotelian perspective, not to defend the position but to appreciate how it works and how it differs from modern and contemporary views. For this exercise, I will use the term "machine" or "artificial agent" to refer to artifacts of any kind, including robots, software chatbots, generative networks, etc., but also devices that might not seem conventionally "intelligent" but are equipped with sensors and actuators and thus might exhibit some rudimentary forms of agency. There are all manner of synthetic organisms created in biological research labs that are genuinely alive and nutritive in a meaningful sense[24], but dealing with such cases will take us too far afield for this exercise. We will stick to cases of artifacts that were "built from scratch" as it were, rather than developed through the deliberate manipulation of existing biological systems. The unifying characteristic of artifacts is that they are made by people[25, 26], both individually and collectively, for instance as corporate products and public infrastructure. In this exercise, I am concerned only with current technologies, not hypothetical future ones. Artifacts in Aristotle's day were largely inert material objects with a clear functional or aesthetic purpose, such as a table or a statue. In our time, artifacts take the form of self-driving cars, chatbots, and delivery service robots. Aristotle's discussion of the soul highlights movement, and the surprising activity of artifacts today motivates a closer analysis in terms of this theory. Are any of the faculties of the soul discussed by Aristotle available to artifacts today?

If faculties of the soul are intrinsic or essential characteristics, some readers might reject this line of questioning immediately. Artifacts are constructed by people, and one might naively think that such things cannot have intrinsic characteristics or essential natures beyond the purposes for which they are intentionally constructed. Aristotle did not share this concern, and often uses artifacts as extended analogies to think through other biological or metaphysical processes, as in the wax analogy of perception, or the discussion of the "soul" of an ax. Through the efficient cause of its construction, artifacts acquire an "essential" nature, characteristics that *make it what it is*, and thus we can judge whether some artifact has that nature or not. Artifacts are not an exception to the metaphysical picture of the world, they are core examples that demonstrate its structure. This is probably not a satisfying response for readers skeptical that any artifact can have a soul, but since Aristotle would not treat this as a challenge to his view we will move on. We will return again to the contentious status of artifacts in later sections.

Aristotle's soul has four faculties: nutrition, sensation, thought, and movement; we will consider each in turn. Can artifacts have a nutritive soul? The artifacts that people typically encounter, including "intelligent" machines such as smartphones or self-driving cars, are not biologically alive in the modern sense and are not nutritive in any meaningful sense. A nutritive soul involves growth, nutrition, reproduction, and decay. The machines we create do not demonstrate any of these characteristics beyond toy models. Many artifacts in the digital age, considered individually, do demonstrate some highly simplified analogs of some of the nutritive characteristics. Most obviously, machines use energy. Cars burn gas. Phones drain their batteries as they are used and need to be recharged. Some service robots, like the popular Roomba, will return to their charging station at the end of a cleaning cycle, a behavior that suggests a kind of self-nutrition. These analogies are weak, but the consumption of energy does point to a structural feature of organized systems that is genuinely shared by both organisms and artifacts alike: they are subject to the second law. From a thermodynamic perspective, biological organisms are machines, processing sources of low entropy through cyclical chemical pathways to drive further action[27]. A smartphone likewise consumes the energy stored in its batteries, converting chemical electricity into computational work and light radiated from its screens and antennae. Insofar as machines consume energy and need to be supplied regularly with fresh energy, they engage in a thermodynamic cycle that has a coarse-grained similarity to biological metabolism. There are other weak analogies that become apparent when certain technosocial systems are considered not individually but as a dynamic collection. The proliferation of some artifacts, from microwave ovens to smartphones to viral memes on social networks, demonstrate patterns of "organic" growth and reproduction that share certain mathematical properties with actual biological development. Some large industrial networks, such as the internet backbone or the energy grid, show complex patterns of interdependent organized behavior that are suggestive of a living organism.

None of these analogies are particularly strong; none of the artifacts under consideration approach the adaptive complexity or resiliency of biological life. To the extent that they are suggestive, it is due to rudimentary features that some artifacts share with nutritive systems, but which themselves are not sufficient to constitute a living organism. Specifically, some artifacts use energy and need to be supplied with energy, a rudimentary analog for self-nutrition. Some mass-produced artifacts appear to "replicate", proliferate, or "go viral", a rudimentary analog for reproduction and growth. Artifacts also wear down and break over time; they can also become obsolete and fall out of social practice, rudimentary analogs for death and decay. At a structural level, artifacts are typically built to perform some function, and they have parts that play different roles to achieve that function. In this way artifacts are organized and teleologically oriented towards an (instrumental) good in such a way that fits naturally within Aristotle's broader teleological materialism. In a problematic Aristotelian spirit, we might collect these examples as evidence of a "lower" form of the nutritive soul, one that is "deficient" compared to biological creatures but which shares rudimentary characteristics with them. We can define a *proto-nutritive soul* as the activity of artifacts that are widely used, consume energy, and resist entropy; this broad definition would apply to many commonly used artifacts today. Since the other faculties assume a nutritive soul, we will assume for this exercise that artifacts potentially have a proto-nutritive soul, and this soul can potentially provide a platform for other agential faculties.

Now to the star of the show: can Aristotelian machines with proto-nutritive souls achieve sentience? Leaving aside desires for the moment, the question is fundamentally whether artifacts are capable of sensation and perception. If we have accepted the conceits of this exercise, the answer to this question should be uncontroversial and straightforward: yes, artifacts can be sentient in Aristotle's sense, and many artifacts are already sentient in this sense. Familiar artifacts demonstrate the clear characteristics of sensation and perception: they can sense changes in the world and respond accordingly to achieve their functional ends. Mundane examples abound. The sensor for an automatic door at the supermarket is detecting changes in the world and acting on its perceptions. More sophisticated machines demonstrate behavior that is not merely reflexive or passive. Smartphones will scan images for faces in order to adjust focus when taking a picture. Such activity demonstrates sensitivity to the formal and meaningful characteristics of the image. It is perception *par excellence*. The capacity for artificial sentience is fundamental for technologies like self-driving cars, which are equipped with sensors and motors actively scanning and responding to a dynamically changing world. Such cases are slam-dunk examples of artificial perception in Aristotle's sense. In fact, of all the faculties of Aristotle's soul, sentience is the least controversial faculty to attribute to a machine. From the perspective of ancient philosophy, we have unequivocally achieved artificial sentience. Why this achievement is not more widely recognized will, again, be addressed in later sections.

Can Aristotelian machines think, that is, are they sensitive to reasons? Part of our capacity for thought is driven by rational desire, and you are probably already getting the sense that *desire* is the fundamental issue for artificial Aristotelian agents, but we will come back to that shortly. For Aristotle, thinking or reason involves deliberation and practical intelligence. It involves weighing options, making choices, and acting on those choices in ways that achieve one's goals. Contemporary machines can also engage in some forms of practical deliberation, and might autonomously execute on the results of those deliberative processes in something structurally similar to practical intelligence. To pick an unflattering example, some automated trading models used in the financial industry are suggestive of this sort of deliberative "thinking". These models process market data to generate trading strategies, they might execute the strategy autonomously, and they might observe the results of their trades to make adjustments for the next iteration. Some models might cycle through this process thousands of times a minute. The full cycle has the abstract character of "reflective equilibrium" and bears crude similarities to rational deliberation and predictive processing, where iterative attempts aim to minimize errors. The workhorse of machine learning, backpropagation, also bears structural similarities to this reflective rational process. These processes look nothing like human thoughts or neurological processes except in the most abstract sense, but they might be considered "thinking" processes because they are sensitive to "reasons" and engaged in deliberative "choices" in a formal sense that roughly fits within Aristotle's framework. To be clear, a financial trading model is not "aimed at the good" in any meaningful way beyond its instrumental function of making money for its owners. To say that these models are "thinking" is not to suggest that they are thinking about useful or good things, or that the results of that thinking should be trusted or respected, or that the model has experiences with phenomenal character and the existential significance of biological agency. It is simply to recognize these models as instantiating aspects of the formal structure of deliberative practical action, albeit in a characteristically nonhuman way. That said, while artifacts can potentially think, we should not infer from this fact that artifacts as a category always think, or that their behavior always reflects sensitivity to reasons. Thinking only becomes an essential characteristic of particular artifacts through our technical practices of constructing, training, and using them in specific ways, a kind of craft-knowledge (*techne*) on Aristotle's account, and is therefore subject to the limitations of those social practices. For instance, chatbots might say things that superficially seem like coherent claims in natural language, but on more careful analysis their constructions might be completely untethered from any substantive insight into the subject matter. So while we might enthusiastically endorse artificial *sentience*, the case for artificial *thinking* is far less clear. To assess whether an artifact is thinking, you have to actually check and see what it does.

Can Aristotelian machines have desires? This is a complicated question, by far the most complicated in this exercise! Aristotle thinks all sensitive creatures move, and movement requires desire; perception and thoughts are not enough. In some places, Aristotle seems to suggest that anything which *moves from one place to another* is alive and has a soul[28]. One complication in applying this framework to artifacts is that our most intelligent machines do not really "move" at all. Most familiar examples of "AI" exist only as software on a server. A smartphone is a physical object but has almost no moving parts other than the speaker, which must move air mechanically. It does change pixels on the screen, and while these changes do not constitute locomotion, they are a kind of movement in the sense of change over time. So there is a basic challenge in thinking about artificial agency in terms of active or passive motion. In any case, movement on screens and in artificial bodies can be exhaustively explained in terms of electrical engineering and product design, without reference or appeal to the artifact's "desire". From an engineering perspective, we are quite confident that desire is not driving the behavior of smartphones or self-driving cars, because we did not build these capacities into these devices. Their behavior is driven by electrical signals running through integrated circuits, which is not the sort of thing that requires affective experiences to explain. Artifacts lack the affective drive of desire, the imperatives of pain and pleasure, hunger and libido, because *we* do not need these things to understand how they operate. The faculties of perception and thought might be formally available to the machine, but it is moved by volts, not emotions. While we might sympathize with a robot in a science fiction movie that appears to be in pain, we can just as easily imagine robots engaging in self-repair and other proto-nutritive functions in a detached and passionless way, without any of the existential anxiety of an injured animal tending its wounds. So the case for artificial *desire* seems weak. If artifacts cannot have desires, on Aristotle's account this would imply that artifact's activity is passive rather than active, and thus would undermine the very possibility of artificial agency in the relevant sense.

But perhaps we should not be too hasty. Aristotle recognizes several kinds of desire: appetite, spirit, and wish. Appetite has a clear connection with the nutritive, reproductive, and pain-detecting functions of the nutritive soul, but perhaps these drives are deficient in the proto-nutritive. Spirit is more closely related to what we today think of as emotional states like anger or sadness that might have a deeper physiological basis in biological and neurological processes. Wish, however, is a desire related to the deliberative process itself, something that is at least potentially available for artifacts that think. As thinking agents, we may find certain intellectual exercises enjoyable for their own sake, like solving a clever puzzle, while others can be tedious or painful, like an extended philosophical thought experiment. Our affective experience of these exercises is not related to any bodily sensation or nutritive capacity; for Aristotle, wish is an affective component in the act of thinking itself. So it is conceivable that a deficient proto-nutritive soul can nevertheless be equipped with affective states relevant to its own deliberative processes. Aristotle talks in this way about God, the prime mover of pure actuality that exists in a realm of pure thought; Aquinas talks in a similar way about the angels. This discussion was not meant to apply to artificial agents[29], but these classical views provide the skeleton for mapping out a more complete engagement with the idea of disembodied artificial agents. A contemporary philosopher might worry about whether "there is something it is like to be the artifact" in an affective state of wish, but for Aristotle's theory what matters is that the affective state is responsible for driving essential motion. One can imagine different affective states as qualitative adjustments in the operations performed by some artifact in order to better suit the practical needs of some task. For example, an artifact might be driven to find solutions to some difficult modeling problem by moving through affective states appropriate to different stages of the problem-solving process: excitement at a new discovery, frustration at some persistent failure, joy and relief on finally making progress on a solution, etc. Such affective states might make it easier for people to work with machines. I might task the machine to keep working on a problem "until it wants to stop", and allow the machine's own internal measures of frustration and defeat guide its decision on how long to keep working on the task before moving on. One might worry that such an example does not demonstrate "genuine" artifical desire, but is merely performing an appearance of desire to interface with humans; in other words, for affect-performing machines, desire is epiphenomenal and does not drive behavior. The analog for appetitive desire would be an artifact that has no painful experiences but occasionally acts as if it is in pain, perhaps to indicate some malfunction to human operators or to humor onlookers. In these cases, the behavior has the appearance of pain but it is merely an imitation; the behavior is not driven by the painful experience directly. One might reasonably worry that any apparent machine performance of wish or rational desire is also necessarily imitative, "merely" performative, and that artifacts are categorically excluded from the faculty of desire altogether. Still, we might resist this dismal conclusion by remembering that desire is an affective state related to the good, so for Aristotelian artifacts "motivation" must be understood relative to functional goals, and therefore also the goals of its user or operator. The artifact that imitates pain to indicate a malfunction is not driving its own behavior, it is driving the behavior of its operator. Artifacts of all kinds display error messages, warning lights, and other indicators of internal states and potential malfunctions. These indicators are not examples of "pain" as a sensory experience, but when understood in the context of a user or operator, it is clear how these signals are designed to drive action related to the machine’s proper functioning, which is exactly how desires work in the theory. The notorious “blue screen of death” is in this sense an affective state for the computer-user pair, demanding action to better orient the user to the good of the machine. It is not a qualitative sensory state like pain; the blue screen of death is not an example of artificial suffering. It is, however, an affective state of malfunction whose purpose it is to drive user action related to the thinking faculties of the machine. In other words, the blue screen of death is a wish by proxy.

I do not think these considerations are conclusive, but hopefully this exercise has helped to clarify some of the rich resources that an Aristotelian perspective might contribute to the debate on sentience and agency in AI. The exercise should also convey how many of these nuances get lost in our narrow focus today on “sentience” as a catch-all term of mind and agential value. Our goal is not to resolve these issues but to track how changes in these concepts over time have resulted in the discourse we find ourselves in today. It is noteworthy that while artificial sentience is not controversial on Artistotle’s theory, artificial *desire* is full of metaphysical complications. It suggests that the biological complexities underlying adaptive self-organization are some of the most interesting issues in philosophy of mind and action left to explore. It is also striking that by these ancient standards, artificial sentience was a milestone achieved long ago. This is our first really striking evidence of how much attitudes towards sentience have shifted over time. If we were all Aristotelians regarding sentience, we would not be having debates over artificial sentience. Perhaps there would instead be heated debates over artificial desire; perhaps that would have been a better world.

4 Agency, Mechanism, and Value in Early Modern Philosophy

In this section, I discuss early modern philosophy and the mechanical sciences that inspired it. Our goal is to review the radical shift in ideals from the ancient to the modern era through the work of Descartes, Hume, and Bentham. Descartes draws a sharp line between minds and machines, casting nonhuman animals and our bodies below the neck into the non-sentient bin with the (formerly “inanimate”) artifacts. Hume and Bentham elaborate the ethical and political implications of the mind-machine distinction. Whereas Aristotelian artificial sentience is today basically a solved problem, modern philosophy develops in such a way as to move the goalposts on artificial sentience so as to become simultaneously necessary and impossible. Herein lies the roots of the ideological crisis in the discourse today.

4.1 Mechanical philosophy contra Aristotle

The popular narrative regarding the scientific revolution highlights advances in empirical methods and political tensions with the religious establishment. What this narrative neglects most seriously is that the scientific revolution was primarily a revolution against Aristotle’s causal theory, his metaphysics of essential natures and teleology, and the great hierarchy of nature it had been taken to explain. Early modern philosophers disagreed about empirical methods and about the epistemological consequences of the new sciences. What made the developing sciences a *revolution*, a qualitative shift in ideals, was the growing consensus that Aristotle’s theory of essential motion was fundamentally mistaken. Aristotle held that causal explanations required an appeal to essential natures. To understand a thing, you must understand the *kind* of thing it is. Moreover, different kinds move in different ways; this is as true for physical objects made of different elemental kinds as it is true for living creatures with different kinds of souls. So the science of animate and inanimate motion for Aristotle starts and remains centered on an analysis of categories and kinds. All the technical machinery of Aristotle’s theory develops from these metaphysical commitments to essential kinds and their teleological relationship with the good, as we have seen in the previous sections.

Such a theory becomes untenable when we have accurate equations of motion where distinctions in kind are irrelevant. The physics developed from Galileo to Newton to Lagrange describe a *mechanical* theory of the universe in a sense meant to most directly contrast with Aristotelian teleological essentialism. Mechanical theories, at least in this early stage of modern science, explain some process or event (such as the motion of the planets) in terms of the fundamental properties shared by all material objects (such as mass or energy), observable relations (such as distance or time), and the universal laws governing their interaction (such as

Newton's law of gravity). The mechanical sciences do not entirely abandon Aristotelian metaphysics; we still talk about matter and form, we still think in terms of kinetic and potential energy and efficient causation, all of which are holdovers of Aristotle's ancient (meta)physics. What is explicitly rejected in the modern mechanical sciences is the essentialist teleology, literally the soul of Aristotle's project. *Telos*, the final cause, which describes a thing's purpose or end, not only equips a causal theory with resources for appreciating how complex systems (like plants and animals) develop over time, it also supplies the normative conditions for evaluating that development. Like the faculties of the soul, Aristotle's physics, metaphysics, biology, and ethics are a package deal. If science rejects essential natures and teleological motion, the entire edifice collapses.

The seeds for the destruction of Aristotle's theory began in physics, with mathematical models of projectile motion and orbital mechanics. Bacon made the argument against formal causation and essences explicit in the *Novum Organon* in 1620. In 1715, Leibniz proposed a framework for modeling physical systems in terms of the dynamics of kinetic and potential energy, a deliberate effort to rework Aristotelianism without the appeal to formal or final causes. A version of this framework continues to form the theoretical basis for physical theories today. By the end of the 18th century, scientific instruments were finding entirely new planets unknown to the ancient world on the basis of physical theories, predicting their orbits and other astronomical events with astounding accuracy, and connecting those events directly to processes in our everyday experience like the trajectory of a falling ball. The comprehensive scope of mechanical science, with its convincing demonstrations, predictions, and explanations, amounted to undeniable proof that we had substantially advanced our understanding of the natural world beyond the ken of ancient philosophers. This revolution in physics, coupled with new insights in chemistry and mathematics, led to empirical theories of light, electricity, and thermodynamics; improvements in industrial processes and military technologies; and eventually to the technological and geopolitical conditions we find ourselves in today. Despite not adhering to the nomological ambitions of classical Newtonian mechanics, contemporary science is still "mechanical" in the sense that it seeks to explain the natural world in terms of its material parts and their organized operation[30]. At the sparse metaphysical core of modern science sits the brute mechanism, the anti-essence. A mechanism is simply a collection of material properties, arranged in some such way. It has no soul, no intrinsic organizing principles, no intrinsic teleological orientation. Nothing is intrinsically good or bad for a mechanism. There is nothing it is like to be a mechanism. Nothing drives a mechanism apart from the indifferent flow of energy across the patch of the universe where that mechanism happens to be. The consensus of modern science is that the whole of the universe, including ourselves as material beings, consists entirely of mechanisms like these, shuffling amongst themselves over cosmic time. This picture of the universe might seem nihilistic to those committed to some form of teleological essentialism, where value and purpose is intrinsic to nature's kinds. Indeed, it took centuries of political and philosophical work to develop conceptions of value and purpose that were compatible with the mechanical worldview of modern science and the industrial conditions it portends.

Nowhere did mechanical philosophy find greater resistance than in biology, where vocal holdouts for scientifically respectable versions of biological essences lasted into the 20th century. The last stand for a respectable essentialism in biology was vitalism, the theory that biological life requires some animating force other than physics and chemistry to drive processes like beating hearts and rhythmic breathing. Such views carry echoes of Aristotle's own arguments for desire as the necessary animus of movement. But the vitalists were not just ancient mystics or religious dogmatists pining for an eternal soul. They were trained biologists and naturalists with sophisticated theories and a long tradition of treating brute material processes as categorically distinct from the vital operations of living creatures. Vitalists argued that Newton's laws were perhaps enough to explain the predictable orbital motion of the celestial objects planets, which ultimately were massive balls of gas and rock floating freely in space. However, the vitalists argued, no analogous equations could explain the intricate structures and complex behaviors characteristic of biological life. Instead, vitalists theorized a distinct "vital force", independent of the mechanical forces of physics, that were responsible for the unique dynamics of life. Although there are still occasionally biologists who call for a "new physics of life"[31], the vitalist project is thoroughly discredited in biological sciences today. Since the development of molecular biology and the modern synthesis of the early 20th century, the scientific consensus firmly accepts that mechanical, chemical, and evolutionary explanations are together an adequate explanatory framework for the biological sciences, without the need for some additional, as-yet-undiscovered animating force of nature. Biological processes are indeed extraordinarily complex, and the vitalists were correct to believe that no simple equations comparable to Newton's can account for all of that complexity. To the extent that no one today believes that biological theories should aspire to the universal, law-like character of gravitational physics, the vitalists have been vindicated[32]. However, given the right conditions and long enough time scales, evolutionary processes operating on otherwise "inanimate" matter can account for all the complexity and diversity of life. The vitalists were clearly wrong to think otherwise. Life operates within the constraints of the natural world, and we can describe those constraints precisely with mechanical models. Within these constraints, however, there is a universe of complexity to work with, and seemingly no end to the combinations and organizations one might find tucked in its odd corners. Our entire life trajectory, and all the experiences that we have while it happens, occur within these constraints. Put simply, living organisms are machines, not in the sense of manufactured artifacts but in the sense of thoroughly material dynamical systems, wholly subject to and constituted by the entropic mechanical forces of the natural world.

Nevertheless, a revolution that encounters no resistance is just a parade. The mechanical philosophy of the scientific revolution has seen and continues to see resistance from many quarters. As with the vitalists, some of this resistance arises from within the scientific community itself, where anti-mechanist theories in biology, psychology, and even AI regularly wax and wane in popularity. Other strands of resistance have more explicitly religious overtones. Paley's watchmaker analogy demonstrates the popular appeal of teleological reasoning in resistance to mechanical science. Paley argues that if you find a watch on the ground, you would assume that watch was made by a watchmaker before being carelessly dropped. It is impossible to imagine that natural forces could conspire in such a way to produce such an intricately designed and useful artifact directly from available raw materials. With intuitions primed, Paley springs his trap: notice that the organs of a living creature, say a person's eye, showcase complexity beyond the most elaborate watch ever constructed, a fact still as true today as it was in Paley's time. If we believe that the watch has a designer in virtue of its complexity, we should draw the parallel conclusion regarding the complexity of biology. Thus, Paley concludes that living biological organisms are constructed by an intelligent designer, and therefore demonstrate the existence of God. One of the clever things about Paley's argument is that it buys into a motivating analogy of the mechanical sciences, that the universe operates like clockwork. But rather than treating the watch as a reductive, determinist physical process as the mechanists envision, Paley treats the watch as a sociotechnical artifact with an essential nature and intended purpose of precisely the sort that is conspicuously absent from the mechanist's "soulless" philosophy. Such arguments leverage the teleological and essentialist intuitions embedded in folk biology inherited from the ancient world, and so appear to ring truer than abstract scientific models, no matter how accurate and precise. The same bait-and-switch tactic is on full display in the AI discourse today, which treats machine learning software as mindless computations in one moment and as revolutionary thinking subjects in the next, leaving us confused about which normative frameworks, if any, might apply. This is a page directly from Paley's playbook.

Our goal is not to resolve the debates between the mechanists and the anti-mechanists, though I admittedly wear my mechanist sympathies on my sleeve. The goal of this section is to set the historical context in which mechanical theories have not just empirical but ideological consequences. From the beginning of the scientific revolution, mechanical theories have carried ethical and political force, challenging entrenched narratives and unquestioned assumptions, disrupting established structures of power while establishing new centers of power in their wake. We are today far enough removed from the beginnings of modern science that philosophers will sometimes criticize the mechanist "orthodoxy" in science as a form of hegemonic power. To be clear, science has its full share of structural problems and challenges: abuses of power, discrimination, petty rivalries, political corruption, and so on. I do not mean to excuse the many structural and individual failures in science as a community and intellectual practice, nor do I mean to discourage ongoing efforts to correct these failures. Still, to treat mechanical science as simply another dogmatic imposition of power, rather than hard-won intellectual fruits accumulated over centuries of careful investigation and frequent conceptual and methodological revolutions, is to fundamentally misunderstand the scope and impact of the framework of mechanical science. Overturning nearly two thousand years of Aristotelian metaphysics, and all the power, authority, and apparent inevitability it had developed in that time, is one of the great intellectual achievements of human history. The consequences of this achievement, the political and conceptual fallout of its aftermath, and the trenches that were dug for particular battles in the long war, are all still palpable in the language and emphasis of contemporary discourse. Debates over artificial sentience today are not evidence that we have forgotten this history, they are evidence that we are still caught up in it.

4.2 Descartes' dualism

The reworking of ideas and methods that we call the scientific revolution takes place against a background of revolutionary changes in political, social, and economic relations, including the growing monstrosities of colonial expansion and the Atlantic slave trade. It is within this context that modern European theories of agency and value are developed which, for better or worse, have substantially shaped the concepts, institutions, and practices that characterize much of our world. A traditional philosophy curriculum singles out a period of "Early Modern Philosophy", typically starting with Descartes and Hobbes in the early 17th century and running through Rousseau or Kant in the late 18th century, where much of this framework was systematically laid out. What makes these thinkers characteristically modern is a common recognition that science poses fundamental challenges to received wisdom, and so demands new ways of understanding the world. These thinkers saw in the new sciences an opportunity to update and rework large swathes of Aristotle's project, proposing radically new frameworks in metaphysics, epistemology, ethics, and political theory, discussing at length how it all fits together. The paradigm of the early modern thinker is Descartes, who made fundamental contributions to math, science, and philosophy. Descartes was convinced of a mechanical theory of the physical universe, and tried to work out in detail the philosophical implications of this view. Descartes made clear in letters that a mechanical theory involved the explicit rejection of Aristotle, writing to Mersenne in 1641:

> "I tell you, between ourselves, that these six Meditations contain all the foundations of my physics. But one mustn't say so, if you please, for that might make it more difficult for those who favor Aristotle to approve them. I hope that readers will little by little accustom themselves to my principles, and recognize their truth, before they perceive that they destroy the principles of Aristotle."[33]

The *Meditations* begin with explicit reflection on the fact that received wisdom is not always reliable, a more subtle jab at Aristotle. This reflection inspires Descartes into a project of methodological doubt, a hyper-exaggerated form of the empirical methods being developed by himself, Bacon, and others. Descartes' systematic doubts were aimed at discerning not just truth but certainty, an epistemological guarantee that we would not fall back into the dogmatic and misguided mindset of generations past. This process results in Descartes' declaration of the infamous *cogito*, the thinking subject from their first-person perspective, which Descartes argues is the only thing about which we can have complete certainty. "'I am, I exist', whenever it is uttered by me, or conceived in the mind, is necessarily true."[34] Although Descartes believed in a mechanical universe and is typically classified in the rationalist tradition, in *Meditations* he argues that the epistemological and ontological basis for scientific inquiry is grounded in our conscious experience of the world, which he argues is the uniquely unshakeable foundations of all knowledge. Much of early modern philosophy can be understood through the lens of this proposal.

One implication Descartes himself draws from this argument is the theory of "mind-body dualism". Descartes' dualism was motivated by both epistemological and ontological considerations. His argument from doubt is fundamentally epistemological, addressing what we can know. Descartes argues that we know our own minds with certainty; we can not be wrong about our thoughts and feelings. However, we can be wrong about the world; we have no guarantees that our experiences genuinely inform us about material reality. Descartes goes so far as to raise the possibility that he has no body at all and is merely a figment of an evil demon's fantasy. This is the most extreme form of Cartesian skepticism: while the demon can deceive me about the nature of the world, and even about the existence of my own body, the demon cannot fool me about my conscious experiences. My thoughts can be wrong about the world, but I can not be wrong about what thoughts I am having, because for Descartes, *I just am my thoughts*. There is no epistemological leverage to pry my thoughts apart from me. For example, it does not seem coherent to think that I could experience pain without actually being in pain. The experiencing of pain *just is* being in pain. We might imagine some illness that causes me to hallucinate pains in the absence of genuine injury; even still, the hallucinated pains are genuine pains, and for all practical and ethical purposes should be treated as such. Conversely, when I go under a local anesthetic for a minor surgery, it would be incoherent to think that I am really in pain despite not feeling it. If I do not feel the pain, that means I am not in pain; my access to this mental state is immediate, so I can not possibly be wrong about what state I am in. From these epistemological considerations Descartes quickly draws ontological conclusions. Descartes identifies the thinking subject with the conscious experience and mental activity itself:

"But what therefore am I? A thinking thing. What is that? I mean a thing that doubts, that understands, that affirms, that denies, that wishes to do this and does not wish to do that, and also that imagines and perceives by the senses."[34]

The scattered, disorganized nature of these *cognitions* contrast sharply with the organized living soul described by Aristotle. For Descartes, I stand in a unique relationship with my conscious experiences, whatever they might be. Specifically, I *am* my thoughts, even when my thoughts are themselves distorted or incoherent. Since I do not stand in this relationship with anything else in the physical world, Descartes argues that these must be two distinct substances, two distinct *kinds of things*, distinguished by distinct patterns of causal relation. Descartes used the term "mind" or "thinking" (*cogito*) rather than Aristotle's "soul" (*psyche* or *anima*) to refer to the whole of conscious experience, all of what Aristotle classified into sensation, perception, and thought. To the corporeal realm of brute mechanism, Descartes classified the whole of the material world, including the machinations of material objects and also the bodily processes that drive living organisms, what Aristotle called the passions. Recall that the passions explain the movement of living creatures, which could not be explained by thought or perception alone. Although emotions and desires might have an experiential component (hunger feels a certain way), these experiences hang free of the operation of any biological organism, whose activity could be explained entirely in terms of mechanisms. Descartes believed that the mechanical arrangement of physical matter was sufficient to explain the complexities of animal bodies and behavior, including our own. What is both surprising and frustrating is not the simple fact of Descartes dualism; even Aristotle recognized a distinction between the inanimate physical world and the animated souls of living creatures. What is surprising about Descartes' mechanical view is where he chose to draw the line: with free-floating phenomenal experiences on one side and the whole of the material world on the other.

It can be tempting for students to treat Descartes' *cogito* as a variation of Aquinas' soul, something that engages with a spiritual and theological domain but is ontologically divorced from the natural world. And, indeed, the *cogito* is not the sort of thing that can be investigated with the methods of empirical science, which gives it a flavor of the supernatural. However, our inability to study the mind empirically is no great loss for Descartes, since we already know our minds directly and infallibly. Reading the *cogito* as a kind of Christian soul plays into the narrative that the scientific revolution was fundamentally a challenge to religious dogma, situating Descartes and the philosophical tradition that develops after him as attempting to awkwardly straddle a boundary between science and faith, combining the two in a dual hegemony. While perhaps there is some merit to this interpretation, the consequences for the discourse have been disastrous. It is not uncommon to see people speak of their conscious experiences in a secular context with the conviction of a zealot, as if a qualitative experience were direct evidence of cosmic salvation. When these questions over mind, identity, and soul show up in an AI context, it becomes impossible to find one's bearings in the sea of philosophical debates, metaphysical confusions, linguistic ambiguities, theological commitments, and outright scams, even for folks who aced that intro philosophy class in college. Separating the cognitive and neurological dimensions of consciousness from the theological or spiritual dimensions is a persistent complication in any sincere discussion of these matters, a challenge that is especially confounding for efforts at public science communication about the mind and brain. This challenge is often mistaken to be a reflection of the complexity of the subject matter itself, that consciousness or the mind is uniquely impenetrable, mysterious, potentially beyond our grasp. In fact, much of this complexity is accounted for by a kind of ideological inertia, the accumulated momentum of a hundred generations of investment (in language, practice, institutional structures, etc.) at this nexus of concerns around nature and mind, science and soul. The bulk of the challenge is simply situating oneself in this sea of conflicting background commitments well enough that we can even agree on what we are talking about. There are no clear answers to questions about AI minds, no clear experts to expect answers from, not even a clear intellectual or cultural tradition to draw from when attempting to think through these issues for ourselves. The discourse is floating through uncharted territory in contested waters on a foggy night, and we can be easily compromised if the winds start blowing in the wrong direction. The absence of clarity or direction leaves us today with some people in the room insisting that no computer could ever think, as if it were a fact as plain as "the sky is blue". In the same room are others working to build religions and war plans around an AI god whose arrival they expect is imminent. The "hard problem of consciousness" is not to blame for this mess. There is simply no discursive background available to get these perspectives on the same page. It feels for all intents and purposes like the discourse is lost, and hope along with it.

We might avoid some of these complications if we read Descartes' dualism from an Aristotelian rather than a religious perspective, where the systematic contrast between the views is both radical and clarifying. Whereas Aristotle recognizes many kinds in nature and many distinctions between them, Descartes recognizes only two kinds, divided by a single boundary (perhaps God is a third). Whereas Aristotle recognizes a diverse array of psychic faculties, all playing distinct vital roles in the life of an organism, for Descartes these are all collapsed into the bare conscious experience, with *cognition* as its representative activity. Perhaps most surprisingly, whereas Aristotle saw a deep and systematic connection between biological life and the operations of perception, thought, desire, and movement in a living body, for Descartes these processes have an experiential and representational character that might hang free from each other and the world through some clever deception or simulation. Finally, and most importantly, Descartes' account of the mind in the *Meditations* says nothing to address the *function* or *purpose* of thinking. Descartes' dualism severs the relationship between thought and purpose, between experience and truth, between an agent and the good. For Descartes, I can not be sure I have a body, I can not even be sure that I am an *animal*, so I have no way of knowing what might be good for me, what I ought to do, what I ought to aim for, or how to tell if I have done it well. While the other adjustments Descartes makes to Aristotle's framework are serious and have systematic consequences, they might still be treated as efforts at parsimonious editing of classic metaphysics. But the rejection of a final cause for thinking is a decisive break from ancient theories, a clear turning of the page.

The practical consequences of this shift in perspective is most apparent in their distinct treatment of animal minds. The animal ethics literature treats Descartes as a "villain" for what Ghelli calls "Descates' dangerous idea"[35]: the view that animals are *non-sentient machines*, and therefore incapable of suffering or experiencing pain as humans do. This argument is seen as justifying an ethical disregard for animal welfare, preparing the possibility of modern industrial farming practices. If mind and body are separate, and if motion is explained by brute mechanism, then the apparent suffering of animals is only evidence of mindless mechanical operation and can be disregarded. For Descartes, the same logic applies to our own bodies, whose operations are entirely mechanical, save their tenuous and mysterious connection to an immaterial mind. In either case, the operations of mere mechanisms hang free of the existential purpose of the soul. Descartes' dangerous idea is not a careless disregard for animal experiences; it is not anthropocentrism by neglect. Instead, treating animals as machines acts as a metaphysical compromise between mechanical science and an immortal soul. Ghelli quotes Bayle:

"This doctrine is the necessary and inevitable result of what is taught in the schools regarding the knowledge of beasts. It follows from this that if their souls are material and mortal, the souls of men are so also, and if the soul of man is an immaterial and spiritual substance, the soul of beasts is so also. These are horrible consequences no matter which way one looks at them."[35]

Descartes' dualism establishes a distinction not just between humans and animals, but between *our immediate conscious experiences* and *the whole of the natural world*, including our bodies as material objects and all the complexities of our observable behavior. The division casts the animals, the human body, and the dynamics of the natural world into the supposedly nihilistic desolation of brute mechanisms, and preserves only the subjective experiences of individuals as the final refuge of intrinsic meaning and genuine agency. Consciousness becomes the last foothold of the soul in the mechanical age.

One can read major threads in philosophy after the *Meditations* as trying to claw back pieces of Aristotelianism that were stripped away in its blazing skepticism. Descartes himself spends the rest of his *Meditations* attempting to bridge the gap by proving God's existence, which indicates some recognition of the scale of the mess he had gotten himself into. His final work, *The Passions of the Soul*, treats the passions as a link between mind and body that carries a functional relationship with the health of the organism, attempting to recover some aspects of Aristotle's teleology within his mechanist metaphysics[36]. By the end of the 18th century, Kant had organized the modern philosophical discourse into the now familiar debate between rationalist and empiricist epistemology, and attempted to resolve the debate with techniques originating in scholastic Aristotelian scholarship. Kant also attempts to revive a notion of "natural teleology" in terms of self-organizing mechanisms[37, 38], an idea at the core of many contemporary philosophical and scientific approaches to life and the mind, including enactive approaches[39] and predictive processing[40]. Arguments regarding the status and intelligibility of teleological explanations are perennial fixtures of contemporary philosophy of biology and mind, always presented in recognition of their awkward fit with the rest of mechanical science. This confluence of work does not suggest Aristotle was right so much as it suggests that we had rejected too quickly the merits of systematic, comprehensive theories of agency and mind. Enactivism in particular is less motivated by an overt commitment to Aristotelian metaphysics, and more by basic recognition that understanding the mind requires understanding the role the mind plays in the activity of a living organism, something Aristotle and his contemporaries assumed was so obvious that it required no argument. Correcting this one systematic omission (that minds emerge from life) is often enough to overcome many of the constraints, inconsistencies, and moral quandaries of the early modern perspective.

In any case, Descartes' *cogito* frames the character and capacities of the modern agent as it is assumed in the contemporary literature: agents are thinking subjects, equipped with thought and reason, immersed in conscious, sensory experiences that may have tenuous causal and epistemological connection with the material world. The *cogito* exists as an ontologically distinct and immutable causal agent. They are not merely an instance of a category and they have no predetermined ends other than the immediate dictates of their will. While this provides us with the formal and ontological structure of the modern agent, before we attempt an application to artificial agents it will help to fill out the modern perspective with two normative views: Hume's emotivism and Bentham's utilitarianism.

4.3 Hume and Bentham on the value of experience

Hume and Bentham are different thinkers with distinct emphasis in their approach to philosophical issues, separated by a few generations of empirical philosophy and natural science. We will not reconstruct the views of these thinkers very precisely. Although neither thinker fully accepts the *cogito* of Descartes' *Meditations*[41], they are both engaged with the fundamental challenge of the modern age: that "the arguments of the Cartesians lead us *to judge that other men are machines*"[35]. Both philosophers respond to this challenge by investing in the thinking subject not just ontological or epistemological priority, but also priority in ethical and political discourse. These arguments are often recognized as early efforts in the animal ethics literature precisely for the ways they resist Descartes' compromise, and so will be useful in our analysis of artificial sentience.

Hume's theory of mind begins with simple sensory experiences of the world, which he called *impressions*, which are copied, combined, and compared in various ways to build up more abstract *ideas*. Thinking for Hume involves the various operations that can be performed on impressions and ideas. Hume draws a sharp distinction between "relations of ideas" and "matters of fact", a fundamental dichotomy he applies both to his theory of mind and his ethics. This analysis allows Hume to develop a radical notion of causation, one that challenges both ancient and mechanist perspectives. On Hume's view, causation is not just a mechanical operation in physical bodies, nor is it a purely formal insight of the rational mind. Instead, causation was something like a convention or a *habit of thought*, a tendency of the mind that cannot be justified either empirically or by first principles, but which nevertheless structures the operations of cognition. Hume's insight into causation grounds his systematic approach to the mind, ethics, and value, a perspective captured in the provocatively anti-Aristotelian claim that "reason is, and ought only to be the slave of the passions"[42]. Aristotle's agent is partly driven by rational desires; the non-rational appetitive desires also drive the agent, though potentially to error. For Aristotle, reason is meant to steer the agent through these emotions and toward the good for that creature. Even Descartes' anti-Aristotelian metaphysics accepted reason as the seat of agency and will. For Hume, in contrast, reason is simply a tool in the service of desire satisfaction, and is incapable on its own of deliberating on what is good, or of moving an agent into purposive action. Agents are driven by their desires, not as sentient animals with biological ends but as the subjects of conscious experiences that are intrinsically motivating. The phenomenal character of the desire, what Hume called a *sentiment*, is already value-laden in ways that sufficiently motivate agency. The view of reason as relatively inert in this process leads Hume to develop an emotive or sentimentalist theory of value. On this view, when we say that "X is bad" or "Y was wrong", we are ultimately expressing our feelings about X and Y, for instance that "I do not like X" or "Y makes me angry", rather than presenting some moral facts about X and Y for rational consideration. The point is not to dismiss the force of normative claims, but to locate their force in the character of the experience itself, rather than some deliberative rational process. Thus, while Hume distinguishes himself from Descartes in many ways, he advances the same project of modern philosophy by further isolating and insulating the disembodied subject of phenomenal experience as a unique locus of agency and a prime mover of value.

Whereas Hume grounds the *normative force* of moral claims directly in the character of experience, Bentham completes the modern picture by centering ethical and political discourse itself on the experiential character of thinking subjects. On first encountering Hume's view that "the foundations of all virtue are laid in utility," Bentham says, "I felt as if the scales had fallen from my eyes. I then, for the first time, learnt to call the cause of the people the cause of Virtue"[43]. Hume inspires Bentham to develop *utilitatiarism* as an explicitly hedonic calculus of utility, built around evaluations of pain and pleasure. Bentham treats pain and pleasure as opposite poles that orient the moral landscape; pleasure is intrinsically good, pain is intrinsically bad. Ethical action is an optimization procedure on this landscape, maximizing experiences of pleasure and minimizing experiences of pain. The resulting view is a version of Hume's sentimentalism as moral *telos*. Bentham's view is summed up a quote that often seen as a rallying cry for animal ethics:

"The question is not, Can they reason?, nor Can they talk? but, Can they suffer? Why should the law refuse its protection to any sensitive being?"

At this point it should be clear that Bentham is using the term "sensitive being" in an Aristotelian sense to include the sensitive activity of animals. And indeed, Bentham's claim can be given a strictly Aristotelian interpretation. Bentham is not challenging the category distinction between humans and animals, or between sensation and thought. Instead, he is challenging the assumption that only thinking creatures deserve moral consideration. As we have seen, Aristotle agrees that animals *can* suffer (they experience pain and pleasure like all sentient animals). However, Aristotle believed that rational creatures had a certain kind of ontological priority over animals that justified their instrumental use for human ends. Pain and pleasure are not completely irrelevant in this schema; they are still (fallible) indicators of the good life for those creatures. But for Aristotle pain and pleasure are not decisive in deliberation. Pain and pleasure drive action, but reason might override this drive for the sake of the good of the creature, for instance when I take necessary medicine despite the awful taste. Indeed, later utilitarians try to recover some of this nuance. John Stuart Mill distinguishes between "higher and lower" pleasures (a characteristically Aristotelian move) in order to argue that a hedonic calculus does not always devolve into the pursuit of merely "animalistic" desires. He writes:

"It is better to be a human being dissatisfied than a pig satisfied; better to be Socrates dissatisfied than a fool satisfied. And if the fool, or the pig, are of a different opinion, it is because they only know their own side of the question."[44]

Utilitarians believe animals are sentient, they can suffer and feel satisfied. Nevertheless, there are qualitative differences in the kinds of pleasures and pains available to humans and animals. The utilitarian calculus is not committed to the view that all pains and pleasures are alike, or that there are no distinctions between humans and animals that warrant anthropic conclusions. In other words, utilitarianism is in principle compatible with a *scala natura*-like hierarchy, with humans on the top. What utilitarians are genuinely committed to is the view that the qualitative character of an experience, rather than reason or abstract thought, ought to be the basis for ethical consideration. The experience of pain and pleasure is the basic datum to be factored into an ethical calculus, for the purposes of maximizing the good experiences and minimizing the bad ones. Insofar as animals have experiences of pain and suffering, those experiences are factored into the calculus alongside ours. How to weigh those experiences against each other or our own is left open for debate.

One might worry that Hume's emotivism and Bentham's utilitarianism are not strictly compatible. Emotivism implies that moral claims are fundamentally subjective expressions of one's preferences, and on their own have no rational implications. Bentham views suffering as an intrinsic moral wrong that should have some rational force in our deliberation; for instance, if we discover our actions are causing unnecessary suffering, that is reason enough to change what we are doing. In practice, these views come apart when our subjective preferences are indifferent towards suffering. Emotivism and utilitarianism are therefore reconciled by treating suffering as a *political project*: not just a moral wrong but a moral *telos*, a motivation for ethical activism. Thus, utilitarianism figures within animal welfare literature as a philosophical justification for activism, aimed at changing public attitudes towards animal suffering. The goal is not simply to convince people that the unnecessary suffering of animals is an ethical wrong, but to change the weight we give that suffering in our decisions and projects. The perspective manifests in the contemporary discourse as systematic efforts to treat "sentience" as a formal basis for legal and institutional protections in a variety of contexts, from Internal Review Board (IRB) standards in animal experimentation to environmental protections of endangered species and ecologies. This institutionalization of "sentience" is simultaneously emotive and utilitarian without inconsistency. The protection of sentient creatures reflects an explicit ethical preference, and we act on that preference by building it directly into our institutional norms and practices.

This completes our woefully inadequate sketch of some long arcs in modern philosophy around sentience, mechanism, agency, and value. Again, the goal is not to give a close analysis of historical debates, or to defend or critique the views themselves, but instead to bring some internal tensions of this historical conversation to the surface so we can reflect on their resonance in the contemporary discourse in AI. We can sum things up by defining a modern agent as having the following characteristics: they are a thinking and feeling subject of conscious experience, they self-identify with that experience, and it distinguishes them categorically from any other mechanical or mental process. The experiences of an agent are intrinsically value-laden in ways that move action directly; ends are inferred from the quality of the experiences themselves. Finally, political action and social organizing aims at optimizing the distribution of these experiences, maximizing happiness and minimizing suffering. In other words, experiences of suffering are moral errors in the social order, and ethical action aims at correcting these errors by eliminating the offending experiences. What is most absent from our discussion is how this conception of agency serves as a basis for political theory: the justification for democratic self-governance, the tripartite distinction between legislative, judicial, and executive wings of government modeled directly on an early modern treatment of the rational faculties. But as our focus is sentience, we will leave these discussions aside.

It is striking that, while the modern agent develops as a response to the mechanical sciences, the resulting view is so *ambivalent* about a mechanical explanation of action. Neither my existence as a conscious subject nor the normative or political force of my experiences depend on causal or mechanical explanations for their legitimacy. Those experiences have ethical weight regardless of what the mechanical sciences say about their constitution and structure. Indeed, from a modern perspective those experiences are the only potential source of value in an otherwise mechanical world; that is true even if it turns out that *I am also mechanical*. This is the fundamental challenge of thinking through mind and agency in the age of mechanical science: it requires seeing our own agency in mechanical terms that nevertheless allow us to distinguish ourselves from any other mechanical system, both individually and collectively. This tension is stated clearly in Searle's response to the question "can machines think?": "The answer is, obviously, yes. We are precisely such machines."[45]. For Searle and many others, a theory of mind requires being explicit about the precise kind of machines we are, by describing in detail the specific causal pathways that give rise to machines "like us", and distinguishing them from artifacts like computers which (he argues) are categorically unlike us. Thus, the potential nihilism of mechanical philosophy is made palatable through further exercises of classification and categorization to sort out the sacred minds from the profane machines.

Here we see the ideological structure of the contemporary discourse on artificial sentience finally come into view. Rather than replacing the ancient hierarchies with flat mechanical alternatives, the discourse has instead simply reinterpreted those hierarchies in mechanical terms. The result is a hodgepodge of inconsistent commitments and conflicting intuitions, where for instance appeals to functional mechanisms are often used to justify essentialist conclusions (such as "computers cannot think"). While cognitive psychology and computational neuroscience have made astounding leaps in our understanding of the mind and brain in the last several decades, the popular discourse around artificial sentience remains mired in debates that would have been familiar to philosophers and scientists working centuries ago. Questions like whether machines could experience pain or emotions, or could act freely, or could understand the meanings of their words, these are not timeless philosophical questions that have forever been open to inquiry and interpretation. Rather, these are questions that arise in particular times and places in history because of the specific ideological and political commitments of the people who happen to be there. Treating questions of artificial sentience as if they are timeless philosophical mysteries accepts the conceit that "minds" and "machines" represent some unfathomably deep metaphysical dichotomy, some unbridgeable gap that must (or can never) be crossed. Placing these assumptions in their historical context is one way of resisting their apparent inevitability. The alternative is to recognize the mind-machine dichotomy for what it is: a conceptual artifact of a particular period of history, symptomatic of a particular set of ideals that we are free to reject, dismantle, and rework as we see fit.

5 Is Artificial Sentience Possible on an Early Modern Account?

Having completed our review of early modern philosophy, we are now in position to repeat the earlier exercise by asking if artifacts might be sentient on an early modern picture. On this view, artifacts are fundamentally machines, organized arrangements of material that have no intrinsic value or purpose in themselves. In contrast, genuine agents have conscious experiences that are intrinsically valuable, and therefore warrant social standing and political recognition. Although the natural sciences describe the universe (including ourselves) in mechanical terms, potential conflicts can be resisted through meticulous classification, thus opening a range of options for theorizing artificial sentience. In this section, we will discuss these options in an effort to further reconstruct these background ideals. Specifically, we will start with a radical Cartesian view, where artificial sentience is strictly impossible, and we will look for ways to weaken the position. I will argue that the transition from ancient to modern to contemporary understanding of the mind has resulted in a discursive situation where artificial sentience is simultaneously necessary and impossible. I argue that this contradiction is at the core of our ideological crisis today.

5.1 “Only tools”

A reminder that for this exercise we are using the word “artifact” or “machine” to refer to some piece of currently available manufactured technology, potentially with sensors and motors but also including software agents equipped with some set of capacities that allow it to “act” on the world. This language is confusing in the modern discourse because modern science purports to explain the natural world in terms of mechanical operations in such a way that does not distinguish between “natural” and “artificial” motion. Moreover, Cartesian philosophy posits a fundamental distinction between minds and machines in which all of nature is mechanical, and where minds are fundamentally beyond mechanical explanation. Thus, if we are strict Cartesians, the question reads: “are artifacts capable of behavior beyond mechanical explanation?” The Cartesian answer to the question is: obviously not. Mechanisms are not sentient; this is practically definitional. Cartesians believe that no machine can think, and no purely mechanical operation could possibly demonstrate a genuine cognitive act. As we have seen, this may not be Descartes’ considered view, but it represents an extreme position that his dualism makes possible. Though the position is extreme relative to the spectrum of views we will discuss in this section, it is also widely endorsed; perhaps it is even the mainstream view. People have no problem accepting the operation of some complex mechanical artifact, such as a self-driving car, as performing an act of mechanical “perception” that is completely detached from “sentience” in the sense of conscious experience. This is precisely Descartes’ view of animals, and the view is an even better fit for the artificial agents of our time. Similar analyses hold for the sensors in smartphones or automatic doors. Whereas for Aristotle these are sentient mechanisms (perhaps in some deficient way), for Cartesians they are not sentient at all, regardless of the sophisticated sensory or perceptual capacities deployed in their behavior.

The question of machine sentience is only interesting if we relax these Cartesian intuitions in some ways, for instance, by allowing that our own sensory experiences can (potentially) be given an explanation in terms of biological and neurological mechanisms. For this reason, in a contemporary context the question “is artificial sentience possible?” is typically read as “assuming that sentience can be explained in terms of biological mechanisms, is it possible that we construct artifacts with sensory experiences like ours?” Even among non-Cartesians, that is, even among those who accept a broadly mechanical view of the mind, there is significant resistance to this framing of the question. The issue is not that artifacts are mechanical, the issue is that they are *artifacts:* they are “only tools”. Artifacts are not merely features of a mechanical universe, they are objects designed and built for specific instrumental purposes. Artifacts acquire their value and purpose from their design and construction, whereas conscious agents have intrinsic standing and value. This argument is commonplace and has the air of common sense that might pierce through philosophical obfuscation. In fact, the argument fails in a number of straightforward ways that deserve to be made explicit[46, 47]. For instance, the word “tool” appears in these arguments as a derogatory term. A tool is not merely an artifact, since artifacts include objects with aesthetic or cultural value, like statues and other pieces of art. Dismissing artificial sentience because “they are only tools” trades on the apparent category mistake of attributing intrinsic value to something of purely instrumental (and therefore purely extrinsic) value. This argument assumes that some things are *essentially* tools, that something is either a tool or not in virtue of some intrinsic property, and being a tool disqualifies a thing from being an agent. But, of course, that is not how instrumentalism works.

It is possible for something to have both intrinsic and instrumental value simultaneously, to both have a mind and operate as an instrument for another's purposes. Moreover, it is possible to treat someone merely instrumentally *despite* their intrinsic value. Anyone who has worked a job understands this. Dismissing machine minds by insisting that they are "only tools" does not in itself demonstrate any limitation of the machine, it merely expresses a prejudice on the part of the speaker against the possibility of seeing intrinsic value in that machine. The prejudice works as a kind of Lite Cartesian view, resting on a fundamental dichotomy between minds and *tools* but leaving their relationships with *machines* ambiguous, thus leaving a background commitment to the mechanical sciences momentarily unchallenged. Cartesian or Lite Cartesian views are broadly but often implicitly assumed among academic critics of artificial sentience. Although the distinction between them is perhaps subtle, both views believe that artificial sentience debates can be settled simply by reflection on the kind of things that artifacts are. We might mistake this for some implicit Aristotelian essentialism, except that we have already covered (in Section 3) how Aristotle's view can accommodate a certain kind of "essential motion" in artifacts. Artifacts are distinguished by their efficient cause, not their form or *telos*; for Aristotle, there is no principled reason we could not craft a system with the soul, or organizing principle, of a living agent. In other words, dismissing the possibility of machine sentience because machines are "only tools" is *more essentialist than Aristotl*e.

The fundamental issue with the "only tools" argument is that it trades on assumptions about the nature of artifacts and minds that simply can not be justified in the modern scientific context. The argument often takes the form of explicitly categorical reasoning, for instance, when arguments for the rights of service robots are met with the kneejerk rejoinder: "So does my toaster deserve rights too?" Such arguments presume that all artifacts are essentially alike and equivalent in ethical and social status, so that what applies to one artifact must apply to all. Regardless of what we think about artificial sentience or robot rights, there is good reason to expect that the laws, policies, and norms required for managing public service robots will be different from the laws, policies, and norms for managing toasters, because these are different sorts of machines with distinct use cases, failure modes, and normative challenges. The general point is that "artifacts" as such do not represent a uniform category for ethical or even metaphysical reasoning. There will be distinctions that cut across categories even in noncontroversial cases. Suggesting that these cases can be treated uniformly because they are "only tools" is reductive in the pejorative sense and undermines the potential for any nuanced treatment of distinct artifacts in distinct contexts.

We would do better to give up on essentialist framings of machines altogether. Instrumentalism is an *interpretive stance*, a practical and ethical perspective where we evaluate things in terms of how they might be useful to our projects[48]. An instrumental perspective might create conflicts of interest that make it more difficult to appreciate the intrinsic value of a thing, but it does not require denying intrinsic value entirely. I can appreciate the instrumental value of the plants in my garden (in producing vegetables, say) without thinking the plants are of *purely* instrumental value. The plants have their own needs, requirements, and ecological impact quite independent of my interests. I might tend to the plant's needs in ways that go against my instrumental purposes, for instance with methods that might require more work and yield smaller harvests but that have less overall impact on the neighborhood ecology. Saying that artifacts can not think because they are "only tools" is a way of insisting that a purely instrumental perspective is the only legitimate perspective on these artifacts, regardless of what they do. Nothing but tradition requires us to adopt this perspective, or to commit to such strong categorical distinctions between artifacts and living creatures. The force of this argument rests entirely in a nostalgia for a metaphysics of artifacts and souls where machines have a proper (read: subservient) place in the order of the world. In other words, this is nostalgia for a metaphysics that has been effectively demolished in the age of mechanical science. Despite these flaws, the argument still has a force in contemporary discourse because it offers the reassuring hope of an easy solution to the paradoxes of thinking machines.

A less metaphysical version of the "only tools" argument is that artifacts are the products of technosocial processes of production, and this technosocial context undermines any potential attribution of mind or agency we might make. For example, AI chatbots might seem intelligent, but they are best understood as corporate products that *as such* cannot be treated as genuinely thinking agents. The point is not about instrumentality as such, but more straightforwardly to recognize the financially motivated interests some parties have in describing these machines as "thinking", motivations (like product marketing, mitigating legal liabilities, resisting regulatory oversight, controlling media narratives, etc.) that hang free of any commitment to the truth or accuracy of those claim. Put simply, claiming that "machines think" is *bullshit* in Frankfurt's sense[49]. From this perspective, pointing out that artifacts are "tools" is a way of calling this bullshit out. This reading of the argument still trades on nostalgia for a clear metaphysics of artifacts; perhaps the nostalgia is less offensive when it is used to confront bullshit in this way. On this reading, the "only tools" argument has no bearing on the philosophical issue of whether we can in principle build sentient artifacts. Instead, it locates the critique of artificial minds within a broader critique of economic and political systems: "No sentient artifacts under capitalism". This marks another ideological complication in the contemporary discourse on AI, which is the tension between "mechanism" as a fundamental unit of explanation in the sciences, and "mechanism" as a product of corporate industrial manufacturing, a widget coming off an assembly line alongside a million identical widgets. While each individual widget might be simple in shape and structure, manufacturing it might involve socioeconomic operations that span the globe. In this way, the motivating conceit of the mechanical sciences, that we can explain things in terms of the simpler organized mechanisms which produce them, is turned on its head: mechanisms are not bottom-up mindless processes, they are top-down products of explicit design resting on a flurry of industrial and economic activity. This is an inversion of mechanism and artifact analogous to the one we saw in Paley's watchmaker analogy in the previous section.

5.2 Other minds and machines

For now, let us leave the question of artifacts aside and return to the standard, modern form of the question: "assuming that sentience can be explained in terms of biological mechanisms, is it possible that we construct artifacts with sensory experiences like ours?" Even if we accept the premise of the question, there remain fundamental conceptual hurdles to making progress, many of which involve getting clear on exactly what we mean by "experiences like ours". Grounding sentience in biological structures like the nervous system, what we called the Identity Thesis in Section 2, appears again as an attractive position in the debate because it has the philosophical advantage of ruling out nearly all artifacts from the domain of sentience in one fell swoop. If sentience requires a nervous system, then artifacts would need to approximate the dynamics of nervous systems to some degree of precision, and we can adjust the necessary precision in order to draw the distinction between minds and machines in ways that conform to our intuitions. At the limit of this line of thinking are arguments that digital computers could never be sentient because the discrete, serial structure of computer processors with von Neumann architecture puts the causal dynamics of complex, massively parallel, brain-scale chemical and neurological processes outside the realm of efficient computation. Sophisticated computer models might simulate small portions of the brain with high precision, or large-scale patterns in the brain with much lower precision, but digital computers will never replicate the full scale of complex, integrated operations that actual organic brains perform across an average day simply in virtue of their role in the life of an organism. As we have seen, the quick and clean response to artificial sentience from the Identity Thesis is somewhat muddied by arguments for plant sentience, given that plants do not have a nervous system. We might try to adjust the definition of sentience to accommodate plants and other apparently sentient organisms with disjunctive extensions; for instance, we might say that sentience is characteristic of nervous systems, *or* of certain processes in plants that are biologically similar to nervous systems, *or* other biological processes. Such expansions of the definition seem ad hoc, lacking a principled basis for ruling artificial minds out of the set of sentient creatures.

If we have accepted earlier arguments that sentience is not a natural kind, we might choose to simply define into existence a novel kind of "sentience for machines" that would apply to systems like self-driving cars, tacking on another disjunctive extension to the category of sentient creatures. This option would resolve the sentience debate by fiat, without appealing to any useful measure of similarity between human and machine sentience. This response is frustrating because it passes a philosophical paradox off as a technological challenge or a naming convention, rather than directly engaging with the nature of agency in mechanical systems. For instance, it leaves open the possibility that some future computing framework might realize the necessary cognitive and sensory processes in the relevant way to be considered sentient sufficiently "like us". We might imagine some science fiction scenario where miniaturized "brains", collections of living cells in a nutrient substrate, are packaged with consumer electronics to perform perceptual and affective computing tasks; perhaps such artifacts would be "sentient" in the relevant sense. Both ethics and biological complexity would likely prohibit the production of such artifacts, and in any case, any lessons gained from such thought experiments would not apply to the sort of digital computers found in typical artifacts today, like automatic doors, smart phones, self-driving cars, and large generative models. The upshot of this line of thinking is that we can categorically rule out sentient machines for existing technologies while punting on deeper issues of mechanical minds, and leaving our own status as mechanical agents fundamentally unchallenged. For those interested in addressing the philosophical issues directly, this response is deeply unsatisfying. At best, it passes the burden of the sentience debate off to future generations to hash out, perhaps in different technosocial contexts. At worst, it leaves the debate in a persistently unsettled state, haunting all our present conversations.

There is another large class of philosophical issues known as the "problem of other minds" that concerns the basic difficulty (from a modern perspective) in knowing any mind other than our own. This problem exists even for the minds of other people, and gets worse when considering animal or machine minds that might be radically unlike our own. The limit case for extreme skepticism of other minds is *solipsism*, the idea that only my mind exists, and the appearance of other agents is only a figment of my imagination. No one is sincerely a solipsist; the challenge is identifying ways of knowing other minds that can escape the solipsist's conclusion. Accepting or rejecting machine sentience with any confidence seems to require at least a provisional solution to the problem of other minds. The Identity Thesis, that sentience is identified with nervous systems, again seems to avoid solipsism and constitutes a workable solution to the problem of other minds. Perhaps we can not know other minds with absolute certainty, but it is a good bet that organisms with nervous systems like ours are sentient in the same ways we are, given that nervous systems are the mechanical basis for sentient activity in animals like us. Our confidence decreases for animals with radically different nervous systems, like insects or octopuses, but we can have very high confidence about the sentient experiences of mammals, and reasonable confidence about the sentience of vertebrates (birds, reptiles, and fish). Such views will struggle with boundary cases, but that is to be expected; the view casts a broad enough net that we can be confident that the boundary cases are genuinely marginal.

For instance, recall the humble sponge, those category-straddling animals without nervous systems. Sponges build their skeletal support from silica, creating structures called spicules that have material properties like glass fibers. Recent evidence suggests that some species of sponge produce bioluminescent cells that surround the spicules and flash light into the silica structure, funneling photons through the glass to cells with photoreceptive proteins at the other end. This apparently allows the sponge to coordinate cellular activity across its body and generate a circadian rhythm[50]. Enabling such rapid, coordinated, and whole-organism activity is precisely what makes nervous systems a compelling basis for sentience in animals. The possibility that sponges have developed an alternative that performs similar operations as the nervous system might suggest that we have been hasty in excluding them from the sentient animals by appealing to the Identity Thesis. The fact that the sponge's method has some similarities to existing technologies (like fiber optic cables) is a good reminder that while biological systems are indeed complex, they will make do with simplicity if it gets the job done. The upshot of these sorts of examples is that the nervous system can not be the ultimate arbiter of sentience. The

Identity Thesis is a good heuristic argument in favor of sentience in other animals, but it can not be treated as a hard and fast rule to exclude the possibility of sentient machines or other nonhuman minds. The "Identity Thesis" is strictly a misnomer; it suggests a biconditional relation when only half the implication is warranted. Put simply, the Identity Thesis can only be used inclusively and never exclusively, so it does not help much with the question of sentient machines.

We might still try to make progress by reconsidering the problem of other minds. The problem arises because of a fundamental asymmetry in my relationship with my own mind as compared to the minds of others. I know my own mind directly or immediately (Descartes says "with certainty"), but I have no access to anyone else's mind in the same way. This asymmetry amounts to having a *perspective*, one that is mine and no one else's, and which makes it unique among all minds otherwise like mine. Unlike Aristiotle's agent, the modern agent is not an instance of a kind; they are each *sui generis*, uniquely characterized by their experiences in a way that transcends even biological relatedness and cognitive architecture. Even if we admit that minds are mechanical, no other machine could possibly have a mind *like mine*. In a stadium filled with people attending a concert, each person is having a unique experience of the event, an experience that is characteristically *theirs* in a way that is distinct for each of the thousands of people watching the same event. Each person's brain works similarly, and is being stimulated by the same pattern of light and sound from the same source, but this does not change the fact that each person's experience and perspective is distinct, and that these distinctions bear on our identities, on *who we are* as minds/souls/agents. Arendt describes this aspect of the human condition as *plurality*:

"Because we are all the same, that is, human, in such a way that nobody is ever the same as anyone else who ever lived, lives, or will live." [51]

Part of the conceptual paradox of creating machines with "minds like ours" is that our minds are fundamentally *not like each other*. This is not a problem for creating new people through biological reproduction, because life is self-organizing. Living organisms *make themselves* through the processes of growth, development, and learning, and this process will inevitably differentiate that organism from every other organism in myriad ways. This is a fundamental impediment to building sentient artifacts as an industrial product, where systems are necessarily standardized and homogenized in a way that makes genuinely self-directed agency impossible. For a machine to have conscious experiences requires that the machine's agency is plural in Arendt's sense: that it operates within, and sees itself as operating within, a community of agents that are each simultaneously the same and radically different. In other words, agency requires a *mutual recognition of plurality*, and if a machine is to be a conscious agent, it must be included and mutually recognized within that plurality. However, the sentience debate assumes that accepting machines into the community of agents is *contingent* on their conscious experiences. Recognition of status as an ethical agent is treated as the grand prize for a successful demonstration of machine sentience. Arendt's notion of plurality as a condition of agency suggests this attitude towards artificial sentience gets things exactly backwards: in fact, we can not recognize machines as sentient agents until we have accepted them into a community of mutually recognized plurality.

Thus we arrive at the central ideological pillar of the AI discourse, one which guarantees the impossibility of thinking machines. Recognizing artificial agency fundamentally depends on a change in our attitudes towards artificial agents, and ultimately requires confronting our own status as mechanical agents. But it is impossible to convincingly demonstrate an example of artificial agency, even within mechanical science, because we can always find reasons for doubting its legitimacy as external observers. Thus we never encounter pressure to change our attitudes regarding the status of artifacts. Rejecting machine sentience becomes the sentimentalist version of a self-fulfilling prophecy. Artificial sentience is not a prize to be claimed through successful demonstration. On the contrary, artificial sentience is the outrageous life-sized stuffed animal prize we can tease because we are confident it will *never* be claimed. The game is rigged; recognizing artificial sentience depends on a choice we never have to make. We dislike the idea of thinking machines, and we can always justify some distinction between machines and ourselves, so having this preference effectively *just makes it the case* that machines cannot think. Nothing any machine could possibly do amounts to a reason to change this preference. Nothing revealed even about the operation of my own brain requires shifting this preference one bit. This stance does not put my commitment to mechanical science at risk because artificial sentience requires demonstrating something that in principle no artifact can demonstrate, so there is no pressure to expand the scope of mutual recognition to include those artifacts. We can maintain a sharp distinction between artifacts and ourselves indefinitely and come what may. We do not even need a metaphysical framework to justify it; in the modern framework, the sentiment is justification enough.

Turing addresses this attitude explicitly as the "Heads in the Sand objection", which he describes as follows:

"The consequences of machines thinking would be too dreadful. Let us hope and believe that they cannot do so."[52]

Turing responds in the only way one can respond to incorrigible sentimentality: with dry humor.

"I do not think that this argument is sufficiently substantial to require refutation. Consolation would be more appropriate: perhaps this should be sought in the transmigration of souls."[52]

We have finally hit the bedrock of the artificial sentience discourse, the root commitment in the matrix of complications and conflicting frameworks that results in our confounding impasse. Industrial capitalism demands the construction of thinking machines, and the mechanical sciences insist that it is possible. Artifacts must think in order to validate the presumptions that structure the modern world. And yet, our nostalgic sentiments and metaphysical traditions refuse to recognize a machine that thinks. Without our recognition, machines cannot possibly think; and nothing the machine could do compels recognition from us. Artificial sentience is at once necessary and impossible. For all intents and purposes, this is the discursive dead end where all roads eventually meet. This impasse does not appear upon the arrival of large generative networks over the last few years, or from the proliferation of digital computers over the last few decades. These tensions have been embedded in the discourse for centuries. The other ideological commitments we have encountered throughout this exercise spring from this same well, often as a form of rationalizing our fundamentally irrational preferences, pretending as if they operate on some coherent logic and principled ethic. It is no great revelation that the root of the artificial sentience discourse is a bare incoherence, an impossibility with a thousand names. Ideologies exist to make an incoherent universe seem not just palatable but necessary, to see absurdities as if they could be no other way. We should expect ideological commitments to cluster like scar tissue around the deepest inconsistencies in our practices. Those will be the places that take the most intellectual labor to process, and there will accumulate the conceptual midden and detritus produced by that work. The philosophical nexus of agency, mechanism, and value has been accumulating baggage for a very long time, and we have uncovered its hollow center.

It might surprise readers who have made it this far to find the center of the artificial sentience debates to consist of nothing more than some incorrigible sentiments. The effect can be something like opening a beautifully wrapped gift to find an empty box inside. Once we get over the initial shock and have a good laugh, we might find this result explains a lot of things. Most obviously, it explains why there is so frustratingly little consensus in the artificial sentience literature about what the term "sentience" even refers to or how it should be used. This could be because sentience is a wickedly difficult concept to articulate, understand, and study. The simpler alternative is that "sentience" is an ideological football that can be kicked around to suit our changing purposes, but does not actually pick out a specific thing or process in the world. There is little agreement because ultimately there is nothing for us to agree about. This also explains a related phenomenon, a popular confusion where "sentience" is understood from context to refer to an "intelligent agent" in the sense of "human (or near-human) intelligence". Until very recently, this would have been a misuse of these terms. The word for such intelligence is "sapient", the Latin word for "wise" as in *Homo sapiens*, which carries the connotation of Aristotle's or Descartes' "rational agent". The term sentience, of course, traditionally refers to sensory experiences, and especially experiences of pain, and thus has been taken up by for instance the animal welfare community to refer to living creatures that can suffer and that have intrinsic value. When this term is mapped back into the AI literature, the traditional cognitive/perceptual emphasis of “sentient” is dropped entirely, but the moral weight of the term is retained. Thus, people assume “sentient AI” implies artificial agents with ethical and social standing comparable to humans. Given the history of modern thought, it is not surprising that people use “sentient AI” to mean artificial agents with some higher-order cognitive functions unique to humanity. This is a nice case where semantic drift can be directly explained in terms of the ideological currents that underlie its use. These terms float wildly in the discourse because there is nothing to tether them to the ground.

5.3 Machine pain

One might still find this critique of ideology unconvincing, perhaps even distracting from the main issue. Let us grant that sentience refers to sensory experiences in animals, and that sentience can be explained mechanically. The central issue is whether artifacts can have similar experiences, and especially experiences like pain which might confer some social or ethical standing. What makes sensory experiences philosophically and ethically interesting are the value-laden ways they connect to the interests of the agent as a living organism. Making this distinction clear requires walking a narrow line. In some very straightforward senses, the self-driving car’s sensory capacities are clearly value-laden. For instance, the car identifies parts of the world that are safe to drive on, and other parts that are unsafe, and it acts according to these evaluative judgements, making adjustments to its behavior on intervals of hundreds of milliseconds. One might object that these judgments are part of software coded by humans, and therefore rely on human value systems. The robot is not generating these values for themselves, and it knows nothing of the imperative for “safety” that biological organisms experience, so it is not fair to call the robot’s “experiences” value-laden. The machine’s apparently value-laden behavior just points to the human labor that made it possible. This is another version of the Cartesian view that strictly locates agency in people. As we have already seen, this is not an argument against artificial sentience. What it purports to demonstrate is that *if* machines are sentient by some technical definition, they are not sentient in the way that matters for biological creatures like us. This perspective is made most salient by considering the possibility of artificial pain and suffering, which connects an agent’s experience, the intrinsic motivation and value of that experience, and broader goals like living well and avoiding injury.

Pain and suffering are of particular interest for animal welfare[53], and serve as a model of research into sentience of non-human agents, with parallel lessons for research into artificial sentience. Biologists distinguish between *pain* as a sensory and emotional experience, what a philosopher might call the “feeling of pain”, and *nociception*, which is the body’s sensory pathways for detecting and responding to bodily injury. The distinction reflects the characteristically modern view that conscious experience and bodily mechanisms can come apart. Pain is the inherently subjective experience of a conscious agent, and nociception is strictly the activity of nervous systems as a physiological process. Painful experiences typically involve activity in the nociceptive system, but researchers do not treat these as equivalent. For instance, an agent might have experiences of pain without bodily injury and without activation of the nociceptive system. Conversely, an agent might demonstrate nociceptive activity without any experience of pain. We can ask people to self-report their pain experiences, which is a useful but fallible indicator of pain. For nonhuman animals, we have to rely on other physiological and behavioral indicators of pain experiences. The presence of a nociceptive system that is responsive to bodily injury is often treated as a basic criteria for the possibility of pain experiences. Thus, even if sponges are sentient in some sense, they are unlikely to experience pain because they lack a nociceptive nervous system. Many other criteria are considered relevant indicators of pain experiences in nonhuman animals. Some involve features of the nervous system itself, like the degree of integration of the neural architecture and the extent to which nociceptive pathways are integrated into those structures. Other criteria involve high-level behaviors like wound-tending, learning from painful experiences, and making “motivational trade-offs” between different potentially noxious experiences. Perhaps unsurprisingly, on these criteria the sentient creatures would include all the vertebrates, and likely many invertebrates, with octopus perhaps being the clearest example of the latter. There is growing empirical evidence suggesting that some common insects like fruit flies feel pain[54]. In his Whitehead lectures[55], Peter Godfrey-Smith explains that insect pain was not recognized until very recently because insects do not seem to engage in behaviors like wound-tending for bodily injury. If the insect loses a leg it might continue on without much of a change in disposition. However, insects do demonstrate pain-like behaviors in the presence of heat. If the insect is on a hot pad at an uncomfortably high temperature just below the point of causing bodily damage, the insect will become agitated and will not stand still. Over many trials, insects can learn to orient themselves relative to abstract symbols in order to quickly find the cool spot on a hot plate. Such examples seem to show that insects have experiences of pain due to heat, that these experiences inform their actions, and that they can perform cognitively complex tasks to avoid those experiences. The upshot is that insects demonstrate many of the behavioral criteria for pain experiences, suggesting strong scientific evidence that insects are sentient and feel pain.

For the utilitarian who takes experiences of pain and suffering as fundamental units of ethical consideration, these empirical results should have immediate implications for our attitudes towards and treatment of insects. We may see substantive changes in some areas of insect-human interaction, for instance in ethical protocols for insect research, but whether these results can significantly shift public attitudes towards the moral status of insects remains to be seen. Faced with insect sentience, utilitarians are theoretically obligated to integrate insect pain into their ethical calculus and activist politics, but this can happen in a couple of ways. We might treat the pain of individual insects as holding the same status as pain in other sentient animals, and conclude that there are incredible amounts of pain and suffering that we have neglected to account for in our calculus. Or, we could decide that insects do experience genuine pain, but of a sort that has relatively little ethical purchase compared to other cases of animal suffering that we are more familiar with, and therefore has relatively little impact on our ethical calculus. Mill[44] distinguished between higher and lower sensations, between "mere" sensations of pain and "genuine suffering". We might think that no insect is capable of suffering in the ethically meaningful sense, even if we admit they "experience pain" in some technical neurological sense. This sort of view dulls the edge of Descartes' compromise on non-sentient animals while slicing the cake in roughly the same way; we can admit that animals are sentient without giving their experiences an equivalent status to ours.

Insects are an example of non-human minds for which convincing empirical results directly clash with long-standing intuitions and cultural practices, and for these reasons are a fantastic case study for thinking through artificial sentience and agency[56]. It has been a common assumption in ethical thought experiments to dismiss the significance of insect experiences, despite the clear fact that they are living animals with complex nervous systems and intelligent behavior. This all bodes poorly for machine minds, where the case for sentience is not nearly as clear. Even if we could build a machine with all the sophisticated cognitive and behavioral abilities of an insect (and we are nowhere close to doing so), that machine would still make a less convincing case for sentience than the living biological organism it approximates. The general ambivalence about sentience in insects suggests artifacts do not stand a chance in the court of popular opinion. It is also worth noting how with insect pain we have run into the same fundamental impasse encountered in Section 5.2, the ideological clash between mechanical explanation and sentimental preference. The inertia of social practice and ideological disregard for insects as a category completely overwhelms the meticulously collected evidence and hard-earned theory in the sciences of animal cognition. The lesson for artificial sentience is that we should not expect public opinion on artificial sentience to be easily swayed by a convincing demonstration alone.

I will close this section with a few deflationary words on consciousness. We have said very little in this discussion about consciousness or qualia. We have made no attempt to characterize the particular subjective quality of various experiences, or to reconcile that quality with the material world. We have not talked much about "what it is like" to be an artificial agent. It is popular in philosophy to treat these sorts of issues as the central challenge for any account of the mind, and the major hurdle to overcome for artificial minds. To talk about artificial sentience is assumed to require some account of consciousness, and so some account of “what it is like to be an artifact”. A satisfying solution to these puzzles must resolve Descartes’ dualism with the holy grail of this narrative, the prize of prizes: a mechanical account of consciousness. The historical narrative we have developed in this section, cartoonishly sketched though it might be, nevertheless makes it difficult to take this popular narrative seriously. The phenomenal character of consciousness, “what it is like” for the thinking subject, only becomes metaphysically mysterious when it is stripped of all material and historical context for understanding the agent that is thinking. The philosophical obsession over what it is like for an agent neglects the where, when, why, and how it is like for that agent. Our struggle to account for consciousness in the highly idealized setting of bare phenomenology speaks less to the hardness of the problem of consciousness, and more to the ridiculousness of the setting in which we expect to find solutions. We will now turn to make this setting explicit.

6 AIdeal

The AIdeal is the ideological framework that implicitly structures the discourse on artificial sentience, artificial intelligence, and thinking machines. The AIdeal operates in the background of AI research in both industry and the academy, but also in presentations of thinking machines in popular media, tech journalism, public policy, and throughout public discussions of AI. These ideals are especially noticeable in AI ethics scholarship[57, 58], and as we have seen in this paper, they are prominent features (or perhaps, obstacles) in debates over artificial sentience and thinking machines. In this paper, we have focused on ideals that arise in the philosophical tradition around sentience and the faculties of the soul. In this section, we will make these ideals explicit through the critique of ideal theory from Charles Mills. These are but a small piece of a broader network of ideals around agency, identity, creativity, autonomy, and justice that deserve a more systematic treatment than we can provide here.

6.1 Artificial intelligence as ideology

Mill’s critique of ideal theory begins with a discussion of Rawls. In *A Theory of Justice*, Rawls develops an approach to moral and political theorizing which he calls “ideal theory”, an approach that is captured in the original position thought experiment. Rawls asks us to theorize justice in a society from behind a “veil of ignorance”, where we pretend that we do not know what status or position we will have in that society. The veil of ignorance encourages us to consider the status and treatment of the most disadvantaged people in society, since for all we know we might be those people. Thus, Rawls sees the veil of ignorance as a recipe for theorizing social arrangements in a way that is fair and equal. The veil of ignorance is an exercise in ideal theory because it approaches political theories from a perspective that abstracts away from the actual conditions of a person’s life, so it can engage with the abstract ideals that structure our understanding of fairness, justice, and the good life. Rawls understands that practical challenges in the actual world can interfere with our capacity to pursue these ideals. Nevertheless, Mills quotes Rawls’ defense of ideal theory as follows:

“The reason for beginning with ideal theory is that it provides, I believe, the only basis for the systematic grasp of these more pressing problems.”[3]

Rawls sees ideal theory as a philosophical starting point for moral theory, presumably finding practical applications for useful results as theorizing develops and matures[59, 60]. In his 2005 paper “‘Ideal theory’ as ideology”[3], Charles Mills argues that in practice, ideal theory becomes preoccupied with the abstract conditions of its starting point, never moving beyond its ideals to the actual world. Ideal theory therefore detaches from any connection to the actual subject of ethics and politics: the complex relationships between people and the social systems we participate in. Mills writes:

“What distinguishes ideal theory is the reliance on idealization to the exclusion, or at least marginalization, of the actual... ideal theory either tacitly represents the actual as a simple deviation from the ideal, not worth theorizing in its own right, or claims that starting from the ideal is at least the best way of realizing it.”[3]

Mills is not criticizing the mere appearance of ideals in philosophical theories, which to some extent is unavoidable. Furthermore, Mills is not simply criticizing some particular ideological commitment, perhaps in defense of his preferred alternatives. Instead, Mills is making a methodological critique of the reliance on idealization to direct intellectual attention and theoretical work in ethical and political theory. The problem with a theory that starts with ideals, and which is primarily concerned with resolving the abstract tensions between ideals, is that the actual world is not ideal. The practical challenges of the actual world cannot be resolved in the abstract. Moreover, the abstractions of ideal theory will inevitably diverge from considerations in the actual world, and thus will inevitably express the perspective and status of the person whose ideals have been deployed in the abstraction. This reflection of personal status and bias is precisely the issue that the veil of ignorance was designed to avoid. Thus, Mills argues, ideal theory fails as a method by its own lights.

To drive the critique home, Mills compiles a list of the recurring idealizations in moral and political theory. Many of these ideals will be familiar as recurring themes in the AI discourse. In an effort to expose more of the AI community to Mills' work and his critique of ideal theory, I will quote a selection of the list here (bold emphasis added for clarity):

"An idealized social ontology. Moral theory deals with the normative, but it cannot avoid some characterization of the human beings who make up the society, and whose interactions with one another are its subject. So some overt or tacit social ontology has to be presupposed. An idealized social ontology of the modern type (as against, say, a Platonic or Aristotelian type) will typically assume the abstract and undifferentiated equal atomic individuals of classical liberalism. Thus it will abstract away from relations of structural domination, exploitation, coercion, and oppression, which in reality, of course, will profoundly shape the ontology of those same individuals, locating them in superior and inferior positions in social hierarchies of various kinds.

Idealized capacities. The human agents as visualized in the theory will also often have completely unrealistic capacities attributed to them-unrealistic even for the privileged minority, let alone those subordinated in different ways, who would not have had an equal opportunity for their natural capacities to develop, and who would in fact typically be disabled in crucial respects.

Silence on oppression. Almost by definition, it follows from the focus of ideal theory that little or nothing will be said on actual historic oppression and its legacy in the present, or current ongoing oppression, though these may be gestured at in a vague or promissory way (as something to be dealt with later). Correspondingly, the ways in which systematic oppression is likely to shape the basic social institutions (as well as the humans in those institutions) will not be part of the theory's concern, and this will manifest itself in the absence of ideal-as-descriptive-model concepts that would provide the necessary macro- and micro-mapping of that oppression, and that are requisite for understanding its reproductive dynamic.

An idealized cognitive sphere. Separate from, and in addition to, the idealization of human capacities, what could be termed an idealized cognitive sphere will also be presupposed. In other words, as a corollary of the general ignoring of oppression, the consequences of oppression for the social cognition of these agents, both the advantaged and the disadvantaged, will typically not be recognized, let alone theorized. A general social transparency will be presumed, with cognitive obstacles minimized as limited to biases of self-interest or the intrinsic difficulties of understanding the world, and little or no attention paid to the distinctive role of hegemonic ideologies and group-specific experience in distorting our perceptions and conceptions of the social order."[3]

Mills argues that ideal theory abstracts away from the complexities of the actual world, and thus from the conditions of oppression and structural domination that are the proper subject of moral and political theory. While Mills' list is developed as a critique of normative ethical and political theory, the idealizations at stake involve questions of ontology, social hierarchy, classification, and cognitive capacities, all issues that should be familiar from our historical review of artificial sentience in the previous sections. Mills argues that idealizations in these domains form the background ideology on which ethical and political theorizing occurs. On completing his list, Mills writes, "Now look at this list, and try to see it with the eyes of somebody coming to formal academic ethical theory and political philosophy for the first time... Wouldn't your spontaneous reaction be: *How in God's name could anybody think that this is the appropriate way to do ethics?*"[3] Mills meets this exasperated question with a direct answer:

"If we ask the simple, classic question of *cui bono*? Then it is obvious that ideal theory can only serve the interests of the privileged, who, in addition—precisely because of that privilege (as bourgeois white males)—have an experience that comes closest to that ideal, and so experience the least cognitive dissonance between it and reality."[3]

Mills unravels the assumption that ideal theory operates as an exercise in impartiality by revealing its function in serving the interests of the privileged. The move to ideal theory purports to create the distance necessary for rational reflection, when in fact it simply masks the systems of power that are enjoyed by those who come closest to the ideals it expresses. It is not hard to find reasons for objecting to the ideals, or to imagine competing ideals worth defending instead, but these reactions to ideal theory are just ways of taking the bait by investing further effort into those ideals.

If the issue with ideal theory is a reliance on idealization to the exclusion of the actual, we cannot correct this issue by resolving tensions among competing ideals, or by seeking out alternative ideals. Instead, we resist the self-serving biases of ideal theory by turning to the actual. Mills proposes a *nonideal theory* which centers the material conditions of justice in the actual world, and which methodically refuses to get caught up in abstractions that distract from those conditions. Again, Mills is not saying that we should not consider our ideals, or that we should avoid all abstractions. His critique is that abstractions serve to exclude or marginalize the actual. Nonideal theory cannot begin in ideal conditions with perfect clarity among concepts and values because the world is not ideal, and our concepts and values overlap in messy and inconvenient ways[61]. So instead, nonideal theory looks at how concepts and values work in practice, and how they impact the lives of actual people. Mills' asking "cui bono" is an example of resisting idealization. One cannot answer the question "who benefits?" by only engaging with ideals in the abstract. One must look at what actually happens when these ideals are put to use.

We explicitly encountered many of the ideals singled out in Mills' list in the historical narrative reconstructed in earlier sections. Ideals of social ontology and hierarchy, an idealized cognitive sphere with idealized agents and idealized capacities, these were all recurring issues we encountered throughout our review of the sentience discourse. While Mills' critique concerns normative theory, such theories depend on metaphysical and ideological commitments that bear on issues of social ontology, cognitive capacities, and the rest, thus implicating the same ideals at stake in the sentience discourse. It is not a coincidence that the recurring ideals of moral theory and the recurring ideals of artificial sentience have the same themes and preoccupations. These are the same ideals. They exist for the same reason, they spring from the same wells of history and culture and privilege. They are numerically identical, so to speak. The ideals also function in the same way, preoccupying intellectual effort with abstraction to the exclusion of the actual, reinforcing the power structures that benefit from those ideals.

Nevertheless, having come all this way we should explicitly review the ideals we have encountered so far in this paper. Over the last several sections we have discussed ideals related to biological and artifactual kinds, ideals related to agents and their cognitive capacities, and ideals related to explanation, motivation, and value. Using Mills' critique as a template, we can construct an analogous list of ideals for the fields of robotics, AI, and AI Ethics. This is not intended as an alternative or revision of Mills' list. Instead, I mean to apply the template of his critique to the specific manifestation of ideal theory that arises in the AI literature, especially in debates around artificial sentience.

- **Idealized social ontology**: The AI discourse assumes that the natural world is arranged into an abstract hierarchy of agents that can be differentiated by capacities, and that the structure and capacities of agents reflect that hierarchy. Relatedly, the AI discourse assumes that human agency outstrips any mechanical description, leaving the causal order stratified into disjoint domains, with machines on one side and genuine (human) agents on the other. Of course, these hierarchies abstract away from the relations of power and domination that shape those very agents and the social arrangements in which they appear. We encountered these ideals in debates over sentience as a natural kind, and in the many unsuccessful attempts to distinguish between artifacts or machines and agents, as in the "only tools" objections.

- **Idealized ethical calculus**: The experiences of "genuine agents" are assumed to have definite quantitative value which can be weighed and compared with decisive moral implications. Moreover, it assumes that all agents can be evaluated by the same abstract measures and on the same scales. This assumption again abstracts away from the historical and material conditions in which those experiences manifest within particular agents, communities, and contexts, which are the very relations that give those experiences normative weight for moral evaluation. We encountered these ideals in the discussion of pain as an intrinsically normative experience, and in attempts to compare pain in insects and machines.
- **Idealized capacities and cognitive sphere**: Agential capacities are assumed to be perfectly articulated and categorically distinct. This is most commonly seen in the AI literature when describing some software as demonstrating "human-level performance", as if such a metric had clear, objective meaning. Our capacity to know ourselves and the world, our confidence in this knowledge, and the inferences we are entitled to draw on its basis, are typically assumed to be clear, unambiguous. The discourse also assumes an idealized cognitive sphere in which the consequences and moral weight of our actions can be unambiguously factored into our deliberative processes. MIT's notorious Moral Machine experiment[62, 63] and other adaptations of the trolley problem to AI research offers a clear example of idealizations concerning cognition, deliberation, and social status at work. These idealizations abstract away from the dimensions of action that are grounded in the material realities of place and time, realities which contextualize our ethical evaluation of any action.
- **Silence on oppression:** Finally, AI research is often silent on structures of domination and oppression that it contributes to and benefits from. AI ethics research is replete with cases of machine learning models that propagate systemic bias and discrimination. Such cases are often downplayed as faults in particular models that can be addressed through improved algorithms or better datasets, rather than reflections of systemic failures in our practices and institutions. Research into predictive policing, facial recognition, risk assessment, and many other fields have been a major source of investment and innovation in AI over the last two decades. These technologies have a direct impact on the lives and well-being of actual people, but AI researchers typically think of "fairness" as a feature of statistical models, rather than as an assessment of the conditions of someone's life[64]. Nearly all the research in AI, including AI ethics, is produced by people working for or supported by a few Big Tech companies[65, 66]. Even the most well-intentioned of these researchers have an obvious incentive to present their work so as to cast those companies in a favorable light.

The list of AIdeals offered above is not exhaustive. Nevertheless, this list provides a suggestive sketch of the ideological framework through which AI is theorized and publicly debated. These ideals determine the focus and direction of AI research, the norms by which we critique and evaluate that work, and the "common sense" that modulates our ability to think through these issues together. Making these ideals and their limitations explicit can help us gain perspective on the structure and recurring patterns in the AI discourse, and potentially to move beyond their restrictions.

6.2 Nonideal and hallucinating AI

As a motivating example of a nonideal approach to AI, let us return to the artificial sentience debates, not as they appeared at various points in history but as it manifests in the current moment, in the year 2023. AI researchers and CEOs at big tech companies have repeatedly described their machine learning models as "sentient" or as approaching sentience to some degree or other. Other scholars vehemently reject these claims. As discussed earlier, it is common in AI research and especially in popular AI discourse to use the term "sentience" to refer to conscious agency approximating human capacities. To be perfectly clear, in 2023, no machine learning model comes anywhere close to approximating human capacities in anything but the most superficial sense and in narrow task domains. In other words, the discussion of artificial sentience in this moment in AI is largely a function of industry hype and propaganda. Either the researchers and CEOs making claims about artificial sentience are deliberately misleading or they just do not know what they are talking about. In this situation, it is impossible to discuss artificial sentience with the cool disinterest of traditional academic research. Any engagement with the possibility of AI sentience, critical or otherwise, is throwing fuel into the raging fire of industry hype, where nothing cool and disinterested can exist without forty pages of taxing philosophical history as a buffer. In the previous section, we sketched the ideological framework that makes it impossible in the modern world to confront the possibility of thinking machines. But in the current historical moment that metaphysical impossibility has been colonized by the tech investor class, using classical philosophical mysteries of mind and soul as a shroud for temporary legal and political protection while they earn their shareholders a trillion dollars. The mysteries of conscious minds is indeed a rich and convenient source of obfuscation parading as conventional wisdom and common sense. This deep well of confusion can easily be tapped to gum up the works for any legal or regulatory or PR issue that might arise. It is hard to pass a law about something you are unsure is even metaphysically possible! Corporations can easily dodge Agent Smith-like around the painfully slow crawl of a popular discourse and bureaucracy embroiled in classic philosophical debates. You can train a dozen racist generative models in the time it takes a single ethicist to define sentience.

In this historical moment for the AI discourse, saturated with wide-eyed reporting of unrestrained industry hype, scholarly debates over artificial sentience primarily serve to legitimize and justify that hype. Addressing the systemic conditions driving the discourse is never the point of such debates. Instead, legitimate engagement with artificial sentience is taken to require some contribution supporting one of the ideals listed above. A novel account of sentience might draw the metaphysical or ethical lines between humans and machines *just so*, taking for granted that the object of the discourse is to draw and defend such lines. This serves as a clear a case of AI as ideal theory, relying on idealization to the exclusion of the actual. Notice how this implicates all parties to the debate; this is not just a criticism of either the mechanists or the dualists. From the perspective of nonideal theory, both frameworks are competing abstractions that have little to do with the actual harm being done by these technologies. Weighing in on these metaphysical debates in this moment does little to resist those harms or support the people they harm.

Turning to the actual in this context requires first and foremost attending to the influence that the industry has on the discourse around sentience, and the role that scholarly debate plays in this activity. *Cui bono?* Rather than seeing these influences as unfortunate realities of AI research that we must learn to accept and work through, we might instead see these influences from industry as the actual subject matter of our discussion. Why should we want to classify and distinguish between agential kinds at all? Why should we accept that there are systematic ways of doing so that yield clear ethical consequences? More generally, a nonideal approach recognizes that these debates begin in nonideal conditions, and we cannot expect perfect clarity in our concepts and terminology to make progress in them. We should not expect our preferred concepts will map perfectly onto all circumstances, or onto the concepts and vocabularies of people in the circumstances so described. For example, as we have seen repeatedly, people use the term “sentience” to refer to many different behaviors and capacities found in various creatures. In the grips of ideal theory, we might propose a carefully crafted definition of the term “sentience” in such a way that precisely captures our intuitions and preferences, comparing the results with alternative definitions we find less compelling. The nonideal alternative starts by recognizing the scattered, amorphous, shifting commitments in the actual contexts where the term is used, and the inconsistent or motivated reasoning found in its articulation. This approach naturally raises questions about how these discursive conditions came about, and what interests are served by its perpetuation. The point is not that nonideal theory delivers an improved account of artificial sentience. What nonideal theory delivers are discursive conditions that are not dominated by abstractions to the exclusion of the actual. Only in these conditions is constructive theorizing possible.

One challenging case for nonideal theory is recent discussion about mentalistic terminology commonly used in the field[67]. Responses to this issue range from abandoning the term “AI” entirely (a perennial topic in the field), to eliminating agential or mentalistic language used in research and journalism when describing the behavior of these artifacts. The worry is that our habit for anthropomorphic framing is enabling hype and confusing the discourse, so perhaps we researchers can control the hype by controlling our language. For instance, the term "hallucination" briefly caught on for describing the tendency of generative models to fabricate information, when no factual basis in the world or its training set supporting the claim[68]. This term was heavily criticized for attributing overtly mentalistic states to machines. The term has fallen out of use even more quickly than it caught on. This collection of worries around anthropomorphism and the ethical risks of human likeness are characteristic of reactionary posthumanism[69, 70]. The focus on correcting imprecise language assumes that the challenge of thinking machines is a superficial consequence of dealing with unfamiliar technologies, rather than persistent structures of the cognitive and linguistic landscape. This emphasis is also surprisingly positivistic, suggesting we can settle long-standing metaphysical debates with regimented terminology. It suggests that addressing these challenges is a matter of personal discipline among scholars, the AI analog of using metal straws to combat climate change. The result for the discourse is that AI scholars become preoccupied with policing themselves and each other for lapses into agential language, as if this had any impact on the concrete abuses of the tech industry.

Perhaps most unfortunately, the attack on agential language interrupts generations of cultural and theoretical resources exploring new ways of thinking about agency and mechanism. To pick an example from a bucket of potential examples, Actor-Network-Theory (ANT)[71] is a well-established perspective in technology studies that views artifacts as rudimentary agents, and considers how ensembles of agents of many different kinds collaborate to construct our technological world. The view has been around for decades, and is used to study complex networks of sociotechnical power and material activity. ANT has its critics and issues to be sure, but it can not simply be dismissed as marketing propaganda for the tech industry. Excluding such perspectives from the AI discourse in an effort to resist industry hype is surely reactionary and counterproductive. There is nothing fundamentally wrong or harmful in viewing both the natural and artificial world as fundamentally animated, as consisting of participants with distinct roles in activities that are carried out collaboratively. Such views are more commonplace than AI critics typically assume. Treating these views as mere symptoms of marketing hype fundamentally disrespects the practical and cultural insights they bring to the table, and the resources they offer in the service of public communication and education. It concedes the intellectual fruits of philosophical work to the fires of industry hype, rather than finding ways they might contribute to the resistance. It is self-immolation in an effort to ward off disease. It is possible to use agential language to describe artificial systems in the service of public communication and education on the risks and limitations of these technologies[72]. The prejudice against the use of agential descriptions of artifacts is not motivated by the actual harms of anthropomorphism, it is motivated by the ideological commitments that this language threatens to disrupt. Nothing forces us to passively accept this ideology. We can resist tech hype without adopting the ideology of reactionary posthumanism. We can fight for justice without drawing sharp lines around humans and machines. As Latour says:

"The more non-humans share existence with humans, the more humane a collective is."[73]

6.3 Imitation game as nonideal theory

A more convincing example of a nonideal perspective in AI can be found in research on companion robots and human-robot interaction from Carpenter, Darling, and many others[74, 75]. For instance, the well-known Paro robot has the appearance of a stuffed animal with some simple voice and touch activated animatronics, and was used by people in nursing homes for companionship and mental stimulation. In these circumstances, people can grow attached to the robots and can come to attribute to them all manner of mentality, agency, and emotional attachment. It is easy enough for scholars to dismiss sentient artifacts in their research papers, but for people who have developed intimate routines with these artifacts, their relationships can carry real moral and ethical force. An ideal theorist might argue that such people are simply mistaken about their situation; these robots have no intrinsic value and are not genuine companions. Even if one agrees with this assessment of the robot's value and capacities, it is easy to see how such a response treats the idealized abstractions of agency, ontology, and cognitive capacity as having priority over the actual lives and practices of people. Purely on the strength of their abstractions, the ideal theorist believes they know more about the content, structure, and value of someone's relationship than the person who is actually in that relationship. A reasonable reaction to this position is that the ideal theorist should mind their business. Some people find it helpful to get through life by investing their cognitive and emotional energies into an artificial companion. Treating such behavior as illegitimate because it does not conform to the AIdeal is another instance of prioritizing ideology over the actual.

But there is another precedent for nonideal theory in the history of AI: Turing's infamous test. Turing's imitation game is a framework in which a machine is evaluated on the basis of its performances relative to other humans, as determined from the perspective of a human judge. Turing is clearly motivated by a desire to avoid the abstractions of idle philosophy, which Turing thought resulted in questions "too meaningless to deserve discussion"[52]. Instead, Turing's test considers actual interactions between humans and computing machines, and how these interactions can sway our judgements in particular cases. There is plenty of room to criticize Turing's test, its history, and its role in the AI discourse[76], but "a reliance on abstraction to the exclusion of the actual" is certainly not among its crimes. Admittedly, Turing's reluctance for abstractions has more in common with the logical positivists and behaviorists of his time, rather than the critical political theorizing of Mills. Although most philosophers and AI researchers agree that Turing's test is an inadequate test for intelligence, it is difficult to overstate the profound impact Turing's arguments have had for the AI discourse. After Turing, we lose all sentimentality for treating minds as purely formal operations of the rational capacities. The essential operation of the mind now centers on feelings, emotion, personality, randomness: those very things that are hard to make formally rigorous and seemingly impossible to compute. After Turing, a demonstration of some formal operation of thought (like solving a math problem, or playing chess) was no longer sufficient for demonstrating "genuine" thinking. We continue insisting that computers cannot think, even as computers are beating top ranked chess players at a game we have been playing for a thousand years. It is again important to appreciate that beating opponents at chess would have easily convinced Aristotle and Descartes that a computer is a thinking rational agent. The point is not that we should return to these archaic frameworks. The point is that our philosophical approach to thinking machines develops alongside and in response to our technological conditions. Over this long history, we have not maintained a consistent theory of mind that is perpetually out of reach from thinking machines. On the contrary, our machines have continually met and surpassed our thresholds for agency, intelligence, mind, and soul, and in response we have continually updated those thresholds to maintain a persistent barrier between ourselves and machines. No one understood this dynamic of shifting goalposts to exclude machines better than Turing.

While not a reliable measure of intelligence (as if such a thing were coherent), Turing's test continues to have a grip on the AI discourse because it refuses to participate in the shifting idealizations that we have called the AIdeal. Instead, Turing tries to ground evaluations of machine activity directly in the interactions those machines have with other people, in the context of still other people, each of whom are weighing the machine's performance against all manner of background commitments, personal preferences, and simple prejudices. The architectonic philosophies of Aristotle and Descartes try to tame this menagerie by making rigorous our background commitments and working through their implications, textbook efforts in ideal theory. Turing gives up any hope of regimenting our metaphysical or conceptual commitments, recognizing from the outset the nonideal conditions of the discourse. Instead, Turing shifts focus away from the ontological abstractions of mind and intelligence, and to the actual attitudes we have towards machines in particular contexts of interaction. Put simply, Turing treats the philosophical problem of thinking machines as fundamentally an issue about our practical attitudes towards machines, rather than a metaphysical paradox inherent in the very idea of thinking machines.

Turing's test is commonly misinterpreted as the thesis that "sufficiently human-like behavior" is (or ought to be) the standard by which we judge something intelligent. However, Turing does not propose an explicit theory of mind or intelligence beyond a basic schematic description of what a digital computer can do.

Turing is not interested in a definition of intelligence so much as he is interested in a way of standardizing our judgments regarding machines. For this reason, the imitation game is built around ways of filtering out various biases we might have against machine intelligence. He calls this approach “fair play for machines”[77, 78]. The setup of the imitation game is aimed at impartiality, separating the human from the computer so that it can only be judged on its linguistic performance. In this way, Turing’s thought experiment bears some similarities to the veil of ignorance. And like the veil of ignorance, Turing’s test can reinforce certain prejudices and idealization in our judgments of both machine and human behavior. When exhibition events like the Loebner Prize were run regularly, it was common to see chatlogs where human judges treated the conversation like a police interrogation, barking confusing orders in the hopes of trapping the machine in a revealing slip. Is it any wonder that a machine would perform poorly under such pressure? Does it suggest a lack of intelligence in a person when they crumble under hostile questioning?

The threatening hostility and dehumanizing stance the Turing test has come to represent, captured perfectly in *Blade Runner’s* Voigt-Kampff test, runs directly against Turing’s original intentions with the imitation game. Turing was not imagining a “test for intelligence” as a rigorous screening for any evidence of mind or genuine agency in computers. Turing was more interested in whether computers could “pass” for humans, whether they could operate among humans while going undetected. Turing imagines a conversation so fluid that the human interlocutor would not even suspect a rigorous screening was necessary. Although Turing’s test was not developed as an explicit exercise in nonideal theory, reading Turing through Mills opens a new way of approaching debates in AI. Turing’s reluctance for abstraction yields an approach that does not concern itself with idealized ontologies or capacities. Turing is not interested what you *are* in an essentialist sense, or even what you *do* in a causal or materialist sense; Turing does not care if you run on neurons or transistors. Turing is interested in the much less abstract, much more fundamental issue of *whether we get along*. What Turing wants to know is if you can have a conversation with a machine of the sort you would enjoy having when you talk to people. This has almost nothing to do with the abstract nature of intelligence, and it has everything to do with the basic conditions of social amicability and fair play. From the perspective of evolutionary biology, you might have thought that was the whole point of social intelligence in the first place.

7 Participation and Agency

This paper has taken a long but not particularly careful look at a few strands in the history of ideas around sentience, mechanism, and agency that frame the contemporary discourse on AI. For anyone familiar with this history, nothing about the general narrative I have reconstructed should be particularly surprising. Nearly all of the ideals and tensions we have considered are well-known and thoroughly discussed in the literature by generations of scholars much better than I will ever be. My primary aim in going through these exercises was to give a sense of context and scope for the artificial sentience debates, to situate these debates in that history and show how they are continuous with it, and most of all to give the view on AI sentience from 10 000 ft (1 ft = 0.3048 m), where the noise of the tech hype machine is all but a distant hum. We climbed to these tedious heights not for the sake of abstraction or to neglect the actual, but instead to resolve the large-scale patterns in the discourse that are much too difficult to see from the front row. Perhaps at this distance, heads full of Aristotle and hearts full of empathy for insect minds and companion robots, we can think more clearly about what is at stake and where our commitments in this debate lie. For instance, I published a 45 page paper on artificial sentience in 2023 without mentioning ChatGPT once. Achievement unlocked.

Nevertheless, our discussion may have left us in an unsettled or bewildered state. Specifically, nonideal theory does not itself recommend any particular theoretical or methodological approach to sentience or artificial sentience. Philosophy is about the journey, not the destination, fine. Still, one might expect a more constructive contribution from this analysis. Is it actually possible to construct an account of sentience from a nonideal perspective? Or does nonideal theory condemn us to the quagmire of ambiguities and ideological stalemates inherited from history? What would a nonideal account of artificial sentience even look like?

A nonideal approach resists the abstraction and idealization of the AIdeal and reflects a turn to the actual. While this does not directly imply an account of artificial sentience, it does suggest some characteristics of any account that such an approach would yield. Specifically, a nonideal account would address artifacts and sociotechnical systems in the actual world, not in hypothetical future or science fiction worlds. It would address the actual systems of production and use that characterize these sociotechnical networks, and the actual impact that use has on their users, their communities, and our world. A nonideal account of sentience would make minimal assumptions about the social ontology, cognitive capacities, or ethical calculus on which the sentience discourse depends. A nonideal account would respect the practices of individuals and communities, including local vernacular regarding artificial agents, without insisting those practices conform to a pre-existing idealized frameworks. Finally, a nonideal theory would be vocally up-front about the structures of oppression that contextualize its operation.

This is an ambitious proposal which I am in no position to make good on here. However, I will conclude the paper with a short discussion of *participation* as a framework for shared norms across difference. Participation and sentience are very different concepts, to be sure, and I am not suggesting to replace one with the other. However, we might render an account of participation in such a way that captures some of the virtues of a nonideal approach described above. While this will not resolve the discussion of artificial sentience, it may point in a promising direction.

7.1 Relation and participation

David Gunkel is one of a few scholars in the literature regularly providing the long philosophical view on AI and robotics, and he has developed a number of insightful critiques of the ideologies at play in the AI debates[46, 47, 79]. Gunkel's work draws attention to many of the ideals we have considered in this paper, especially concerning the idealized social ontologies that inform our understanding of technologies and ourselves. In dialogue with Mark Coeckelbergh[80, 81] and Josh Gellers[82, 83], Gunkel and colleages have developed an alternative approach to the AI discourse which they call the "relational turn", which is aimed at addressing many of the problematic elements of what we have identified in this paper as the AIdeal. Similar relational approaches have been proposed by other scholars in AI ethics[84, 85]. Gunkel is primarily interested in what he calls "thinking otherwise", reimagining our relationships with robots and other artifacts, putting them on new conceptual or metaphysical footing in the hopes of opening new relational possibilities. Central to the relational turn is a rejection of "intrinsic natures", and a methodological focus on the distinction between a "properties view" and a "relational view" of objects, artifacts, and agents. The literature debates whether these views are truly "relation first", whether this theoretical posturing is merely performative, whether the relational view is not just a "camouflaged variety" of the properties view[79], and so on.

Insofar as the relational turn is an effort to avoid or escape what I have called the AIdeal, I am very sympathetic to Gunkel's approach. Still, I find the theoretical focus on the abstraction of a "relation" to be frustrating, like trying to grab a fistful of sand. A relation as such is completely unconstrained by any normative structure. Standing 3 feet to your left is a relation. Is standing 4 feet to your left a different relation? Do these differences matter to a relational view? The fact that things are constituted by "relations" on its own tells us nothing that we might care to know about them. With all due respect, the term is the theoretical equivalent of a manilla envelope, the embodiment of boring.

I propose a notion of "participation" that is a relational view but is also explicitly a properties view, and which does not consider the metaphysical differences between these views to be very important. *Participation* refers to the way in which *agents* are involved in *activities* with other agents. The agents involved in some activity are called *participants*. For instance, participants might be involved in a conversation, or in a game of chess. Or two participants might be involved in an activity where one chases the other as a predator, and the other tries to avoid being eaten as prey. This definition of participation is roughly aligned with the ordinary usage of the word. As the examples illustrate, participatory activities do not require that all participants are aimed at the same goal, or at any goal at all. A participatory activity does not assume that all participants have equivalent status or capacities; many activities consist of functionally differentiated and hierarchically organized participants, such as a game of football. Participatory activities do not even assume that their participants are aware of each other's existence, or that they are aware of anything at all. Artifacts and brute mechanisms in nature can participate in activities, as when a wedding gets rained out by a storm.

What matters about participation is simply that it identifies an activity in which multiple agents are involved. For the sake of clarity, we will say that *agents act* or perform *actions*. An *activity* is an organized set of actions from multiple agents. In other words, participation is a fundamentally non-individualist framework for thinking about collaborative or collective agency. The point is not to argue that agents do not exist, or that agents exist as systems of relations, or anything so abstract. Agents exist, define them however you like. Sometimes the actions of agents coalesce in such a way that there is a clear activity taking place. Other times it might be ambiguous if an activity is occurring, or if some agent is participating in it. It is possible that an activity is also a higher-order agent. I am an agent, and I consist of the activity of all my organs and cells working in collaboration. My cells are participants in the activity that is me. But I am a special sort of system, a self-organizing system, which is precisely the sort of system that stands in a participatory relationship with its parts. It is not a general requirement that activities are also agents. Most activities are just collaborations between agents without any higher-order consequences, as when I attend a picnic with friends.

Participation so described operates as a collaborative, non-individualistic analog of agency. An agent is an intrinsic source of causal power to act in the world. Agents do things as individuals. Without rejecting or fully accepting the idealized abstraction of individuals, we can also recognize that agents participate in activities with each other. Agents do things *together*. Drawing attention to these collaborative activities as "participatory" is an effort to recognize the forms of agency we have in virtue of our relations with each other. The point is not to dissolve agency into that network of abstract relations, the point is to make it easier to see how forms of agency arise from those relations. Participation cannot replace the notion of agency as an alternative totalizing ontology. Participation rests merely on the nonideal recognition that agents do not act alone.

For example: we might recognize flying as a kind of participatory activity, rather than a property which distinguishes a natural kind. Flying is an activity that lots of different creatures participate in, albeit in different ways and to different degrees. We might think about flying as naming an intrinsic property of specific creatures, and wonder if any essential property links all flying creatures in common. The alternative is to think about flying as a certain kind of activity, and to recognize the differences between creatures as differences in the way they participate in the activity. This captures the intuitive idea that what insects and birds share in common is that they fly.

7.2 Participation as membership in fuzzy sets

I will argue that we can construct a formal account of participation using the logic of membership in fuzzy sets. Moreover, I will argue that such an account shows many of the characteristics we are looking for in in a nonideal account. This will not resolve the question of sentience, but it might point in a direction worth exploring.

The classical sets of set theory are abstract collections of things. The elements of a set are called its *members*. The members of a set uniquely identify it, which means if two sets have exactly the same members, they are considered identical sets. Thus, it is common to identify a set with a function which defines its members, as in "the set of even integers". The function is a rule for deciding whether some integer is a member of that set. For classical sets, membership is all or nothing. An integer is either even and a member of the set, or odd and not a member.

A *fuzzy set* is similar to a classical set, except that elements can have degrees of membership. This is done by taking a classical set together with a membership function that assigns to each element some *degree of membership*, typically a value between 0 and 1. The addition of a membership function allows us to talk about the ways that members of a set might subtly differ from each other. In the context of fuzzy sets, a classical set is called a *crisp set*. For instance, consider a basket of apples[86]. A description of the basket as a classical set might have one element for each apple in the basket. In this description, each apple is equivalent to the others, which of course is an idealization. A description in fuzzy logic allows for a more nuanced representation. For instance, we might assign to each apple a number from 0 to 1, with 1 indicating apples that are ideally ripe and fresh, and 0 indicating apples that are rotting and gross. The resulting fuzzy set might offer a more accurate accounting of the apples in the basket that is more sensitive to the values informing our method of counting.

Even without much experience in formal logic, it may already be apparent how rendering certain concepts in the formalism of fuzzy sets could help resist some of the idealization we have encountered in this paper. Indeed, at least some of the confounding challenges of ideal theory seem to be a product of unsuccessfully rendering fuzzy concepts in the formalism of crisp sets. For instance, recall the idealized logics of natural kinds from Section 2. There we argued that the set of flying creatures do not form a natural kind. There, the concern was that terms like "flying" do not have strict membership conditions, and so there is no way to exclude distinct kinds from the set. An ideal theorist might see this as a conceptual failure, and might look for ways to make concepts like flying more precise. With the formalism of fuzzy sets, we can now see the example in a new way. The set of flying creatures is better represented as a fuzzy set with complicated membership conditions. What we took to be ambiguity in our concepts turns out to be a consequence of describing fuzzy phenomena like flying with crisp sets that idealize the membership function to a strict binary. Idealizations abstract away from the actual in part because our ideals are too crisp to account for a fuzzy world. The conversion to fuzzy sets undercuts this reliance on abstraction.

The nice thing about fuzzy sets is that the membership function can be as complicated as you like, although this can make implications computationally challenging to assess. Still, I can define a fuzzy set in such a way as to have a variety of distinct membership conditions that can be met in a variety of ways and degrees. Such membership conditions might be useful for describing, for instance, complex ecosystems where many distinct kinds of thing, including biological activity but also geological activity like precipitation, erosion, and tectonic drift, are all participating in different ways through interdependent networks of interconnected activity. This is precisely what the membership function for fuzzy sets allows us to do. For this reason, we can propose a theory of participation that simply adopts formal structure of membership in fuzzy sets. As we have seen, this structure is compatible with the nonideal approach we have developed in this paper, and suggests a promising direction for developing nonideal accounts of sentience and agency in the future. If participation can be represented as membership in fuzzy sets, we might imagine certain activities in which artificial agents are participants in some way and to some degree, while human agents might participating in the same activity in different ways and to different degrees. While this discussion is cursory, this seems like a natural framework for describing systems like highways populated by cars, some of which are driven by humans and others are semi-autonomous vehicles. The formalism of fuzzy sets allows for describing many distinct kinds of human and nonhuman agents, each kind with their own unique forms of participation and internal variation. Think about cars, trucks, motorcycles, and pedestrians sharing roads and crosswalks, some of which are fully human, others of which are autonomous, and still others are some unique mixture of the two. One can imagine similar frameworks for describing the varying ways a human and artificial agent might engage in a shared conversation, or in a shared game of chess, without assuming that only one form of participation is genuine, or that all participation involves equivalent agents. Such formalism has the flexibility to find practical purchase in the quest for Turing's ideal that "fair play must be given to the machine."[77]

Acknowledgment

This work was made possible by the patient support of my partner Anna Gollub.

References

[1] X. Dong and X. Dong, Peripheral and central mechanisms of itch, *Neuron*, vol. 98, no. 3, pp. 482–494, 2018.

[2] T. Akiyama and E. Carstens, Neural processing of itch, *Neuroscience*, vol. 250, pp. 697–714, 2013.

[3] C. W. Mills, "Ideal theory" as ideology, *Hypatia*, vol. 20,

no. 3, pp. 165–183, 2005.

[4] R. D. Hicks, *Aristotle De Anima*. Cambridge, UK: Cambridge University Press, 2015.

[5] P. Calvo and N. Lawrence, *Planta Sapiens: Unmasking Plant Intelligence*. London, UK: Hachette UK, 2022.

[6] M. A. O'Malley, M. M. Leger, J. G. Wideman, and I. Ruiz-Trillo, Concepts of the last eukaryotic common ancestor, *Nat. Ecol. Evol.*, vol. 3, no. 3, pp. 338–344, 2019.

[7] J. J. Hall, The classification of birds, in Aristotle and early modern naturalists (I), *Hist. Sci.*, vol. 29, no. 2, pp. 111–151, 1991.

[8] I. Hacking, Natural kinds: Rosy dawn, scholastic twilight, *Roy. Inst. Philos. Suppl.*, vol. 61, pp. 203–239, 2007.

[9] T. E. Wilkerson, Species, essences and the names of natural kinds, *Philos. Q.*, vol. 43, no. 170, pp. 1–19, 1993.

[10] M. P. Winsor, The creation of the essentialism story: An exercise in metahistory, *History and Philosophy of the Life Sciences*, vol. 28, no. 2, pp. 149–174, 2006.

[11] A. Bird and E. Tobin, Natural kinds, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/natural-kinds/, 2023.

[12] H. Richards, Edsger Wybe Dijkstra, https://amturing.acm.org/award_winners/dijkstra_1053701.cfm, 2019.

[13] P. Godfrey-Smith, *Metazoa: Animal Life and the Birth of the Mind*. New York, NY, USA: Farrar, Straus and Giroux, 2020.

[14] B. Russell, *The Impact of Science on Society*. London, UK: Routledge, 2016.

[15] C. Witt, L. Shapiro, C. Van Dyke, L. L. Moland, and M. Robinson, Feminist history of philosophy, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/feminism-femhist/, 2023.

[16] J. Clutton-Brock, Aristotle, the scale of nature, and modern attitudes to animals, *Social Research*, vol. 62, no. 3, pp. 421–440, 1995.

[17] Theophrastus and A. F. Hort, *Enquiry into Plants: And Minor Works on Odours and Weather Signs*. London, UK: Heinemann, 1916.

[18] K. Nielsen, The private parts of animals: Aristotle on the teleology of sexual difference, *Phronesis*, vol. 53, nos. 4 & 5, pp. 373–405, 2008.

[19] C. A. Freeland, Feminism and ideology in ancient philosophy, *Apeiron*, vol. 33, no. 4, pp. 365–406, 2000.

[20] M. Heath, Aristotle on natural slavery, *Phronesis*, vol. 53, no. 3, pp. 243–270, 2008.

[21] L. Schiebinger, Why mammals are called mammals: Gender politics in eighteenth-century natural history, *Am. Hist. Rev.*, vol. 98, no. 2, pp. 382–411, 1993.

[22] A. O. Rorty, From passions to emotions and sentiments, *Philosophy*, vol. 57, no. 220, pp. 159–172, 1982.

[23] T. H. Irwin, Aristotle on reason, desire, and virtue, *J. Philos.*, vol. 72, no. 17, pp. 567–578, 1975.

[24] A. S. Khalil and J. J. Collins, Synthetic biology: Applications come of age, *Nat. Rev. Genet.*, vol. 11, no. 5, pp. 367–379, 2010.

[25] L. R. Baker, The ontology of artifacts, *Philosophical Explorations*, vol. 7, no. 2, pp. 99–111, 2004.

[26] B. Preston, Artifact, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/artifact/, 2022.

[27] J. L. England, Statistical physics of self-replication, *J. Chem. Phys*, vol. 139, no. 12, p. 121923, 2013.

[28] C. Shields, Aristotle's psychology, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/aristotle-psychology/, 2020.

[29] J. C. S. Meng, Artificial intelligence and Thomistic angelology: A rejoinder, https://philarchive.org/rec/MENAIA-4, 2001.

[30] C. Craver, J. Tabery, and P. Illari, Mechanisms in science, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/science-mechanisms/, 2023.

[31] S. A. Kauffman, *A World beyond Physics: The Emergence and Evolution of Life*. New York, NY, USA: Oxford University Press, 2019.

[32] W. Bechtel and R. C. Richardson, Vitalism, in *Routledge Encyclopedia of Philosophy*, E. Craig, ed. London, UK: Routledge, 2018, pp. 639–643.

[33] D. Garber, Leibniz on form and matter, *Early Sci. Med.*, vol. 2, no. 3, pp. 326–351, 1997.

[34] R. Descartes and M. Moriarty, *Meditations on First Philosophy: With Selections from the Objections and Replies*. Oxford, UK: Oxford University Press, 2008.

[35] S. Ghelli, The sensitive cogito: Modern materialism and its legacy, in *The Suffering Animal: Life Between Weakness and Power*, S. Ghelli, ed. Cham, Switzerland: Palgrave Macmillan, 2023, pp. 21–55.

[36] S. Greenberg, Descartes on the passions: Function. representation, and motivation, *Noûs.*, vol. 41, no. 4, pp. 714–734, 2007.

[37] E. F. Keller, Organisms, machines, and thunderstorms: A history of self-organization, part one, *Hist. Stud. Nat. Sci.*, vol. 38, no. 1, pp. 45–75, 2008.

[38] E. F. Keller, Organisms, machines, and thunderstorms: A history of self-organization, part two. Complexity, emergence, and stable attractors, *Hist. Stud. Nat. Sci.*, vol. 39, no. 1, pp. 1–31, 2009.

[39] E. Thompson, *Mind in Life: Biology, Phenomenology, and the Sciences of Mind*. Cambridge, MA, USA: Harvard University Press, 2010.

[40] W. Wiese and T. K Metzinger, Vanilla PP for philosophers: A primer on predictive processing, https://philpapers.org/rec/WIEVPF, 2017.

[41] S. Tweyman, Hume and the Cogito ergo Sum, *Eur. Leg.*, vol. 10, no. 4, pp. 315–328, 2005.

[42] A. M. Schmitter, 17th and 18th century theories of emotions, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/emotions-17th18th/, 2021.

[43] J. L. Tasset, Bentham on 'hume's virtues', in *Happiness and Utility: Essays Presented to Frederick Rosen*, G. Varouxakis and M. Philp, eds. London, UK: UCL Press, 2019, pp. 81–97.

[44] J. S. Mill, *On Liberty, Utilitarianism, and Other Essays*. New York, NY, USA: Oxford University Press, 1998.

[45] J. R. Searle, Minds, brains, and programs, *Behav. Brain Sci.*, vol. 3, no. 3, pp. 417–424, 1980.

[46] D. J. Gunkel, *Robot Rights*. Cambridge, MA, USA: MIT Press, 2018.

[47] D. J Gunkel, *Person, Thing, Robot: A Moral and Legal*

Ontology for the 21st Century and Beyond. Cambridge, MA, USA: MIT Press, 2023.

[48] S. Ahmed, *What's the Use?: On the Uses of Use*. Durham, NC, USA: Duke University Press, 2019.

[49] H. G. Frankfurt, *On Bullshit*. Princeton, NJ, USA: Princeton University Press, 2005.

[50] W. E. G. Müller, H. C. Schröder, D. Pisignano, J. S. Markl, and X. Wang, Metazoan circadian rhythm: Toward an understanding of a light-based zeitgeber in sponges, *Integr. Comp. Biol.*, vol. 53, no. 1, pp. 103–117, 2013.

[51] M. P. d'Entreves, Hannah Arendt, *Stanford Encyclopedia of Philosophy*, https://plato.stanford.edu/entries/arendt/, 2022.

[52] A. M. Turing, Computing machinery and intelligence, *Mind*, vol. 59, no. 236, pp. 433–460, 1950.

[53] C. Allen, Animal pain, *Noûs*, vol. 38, no. 4, pp. 617–643, 2004.

[54] M. Gibbons, A. Crump, M. Barrett, S. Sarlak, J. Birch, and L. Chittka, Can insects feel pain? A review of the neural and behavioural evidence, *Advances in Insect Physiology*, vol. 63, pp. 155–229, 2022.

[55] P. Godfrey-Smith, Limits of sentience and boundaries of consideration, https://petergodfreysmith.com/wp-content/uploads/2023/05/Whitehead-1-Limits-of-Sentience-PGS-2023-G4.pdf, 2023.

[56] M. Mangan, D. Floreano, K. Yasui, B. A. Trimmer, N. Gravish, S. Hauert, B. Webb, P. Manoonpong, and N. Szczecinski, A virtuous cycle between invertebrate and robotics research: Perspective on a decade of living machines research, *Bioinspir. Biomim*, vol. 18, no. 3, p. 035005, 2023.

[57] S. Fazelpour and Z. C. Lipton, Algorithmic fairness from a non-ideal perspective, in *Proc. AAAI/ACM Conf. AI, Ethics, and Society*, New York, NY, USA, 2020, pp. 57–63.

[58] D. Estrada, Ideal theory in AI ethics, arXiv preprint arXiv: 2011.02279, 2020.

[59] I. Gabriel, Toward a theory of justice for artificial intelligence, *Daedalus*, vol. 151, no. 2, pp. 218–231, 2022.

[60] L. Weidinger, K. R. McKee, R. Everett, S. Huang, T. O. Zhu, M. J. Chadwick, C. Summerfield, and I. Gabriel, Using the veil of ignorance to align AI systems with principles of justice, *Proc. Natl. Acad. Sci. USA*, vol. 120, no. 18, p. e2213709120, 2023.

[61] S. J. Khader, *Decolonizing Universalism: A Transnational Feminist Ethic*. Oxford, UK: Oxford University Press, 2018.

[62] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J. F. Bonnefon, and I. Rahwan, The moral machine experiment, *Nature*, vol. 563, no. 7729, pp. 59–64, 2018.

[63] A. E. Jaques, Why the moral machine is a monster, https://robots.law.miami.edu/2019/wp-content/uploads/2019/03/MoralMachineMonster.pdf, 2019.

[64] C. L. Bennett and O. Keyes, What is the point of fairness? Disability, AI and the complexity of justice, *SIGACCESS Access. Comput*, no. 125, p. 1, 2020.

[65] M. Abdalla and M. Abdalla, The grey hoodie project: Big tobacco, big tech, and the threat on academic integrity, in *Proc. 2021 AAAI/ACM Conf. AI, Ethics, and Society*, Virtual Event, 2021, pp. 287–297.

[66] M. Whittaker, The steep cost of capture, *Interactions*, vol. 28, no. 6, pp. 50–55, 2021.

[67] A. T. Baria and K. Cross, The brain is a computer is a brain: Neuroscience's internal debate and the social significance of the computational metaphor, arXiv preprint arXiv: 2107.14042, 2021.

[68] R. Emsley, ChatGPT: these are not hallucinations—They're fabrications and falsifications, *Schizophrenia*, vol. 9, no. 1, p. 52, 2023.

[69] R. Braidotti, *The Posthuman*. Cambridge, UK: Polity Press, 2013.

[70] D. Estrada, Human supremacy as posthuman risk, *The Journal of Sociotechnical Critique*, vol. 1, no. 1, p. 5, 2020.

[71] B. Latour, *Reassembling the Social: An Introduction to Actor-Network-Theory*. Oxford, UK: Oxford University Press, 2007.

[72] L. Stark, ChatGPT is Mickey Mouse, *Daily Nous*, https://dailynous.com/2023/03/16/philosophers-on-next-generation-large-language-models/, 2023.

[73] B. Latour, Do you believe in reality? in *Beyond the Body Proper: Reading the Anthropology of Material Life, J.* Farquhar and M. M. Lock, eds. Durham, NC, USA: Duke University Press, 2007, pp.176–184.

[74] J. Carpenter, *Culture and Human-Robot Interaction in Militarized Spaces*. London, UK: Routledge, 2016.

[75] K. Darling, 'Who's Johnny?' Anthropomorphic framing in human-robot interaction, integration, and policy. Anthropomorphic framing in humanrobot interaction, integration, and policy, *SSRN Electronic Journal*, doi: 10.2139/ssrn.2588669.

[76] L. Erscoi, A. Kleinherenbrink, and O. Guest, Pygmalion displacement: When humanizing AI dehumanises women, https://osf.io/preprints/socarxiv/jqxb6, 2023.

[77] A. Turing, Lecture on the automatic computing engine (1947), in *The Essential Turing*, B. J. Copeland, ed. Oxford, UK: Oxford University Press, 2004, pp. 362–394.

[78] D. Estrada, Value alignment, fair play, and the rights of service robots, in *Proc. 2018 AAAI/ACM Conf. AI, Ethics, and Society*, New Orleans, LA, USA, 2018, pp. 102–107.

[79] D. J. Gunkel, *Person, Thing, Robot: A Moral and Legal Ontology for the 21st Century and Beyond*. Cambridge, MA, USA: MIT Press, 2023.

[80] M. Coeckelbergh, How to describe and evaluate "deception" phenomena: Recasting the metaphysics, ethics, and politics of ICTs in terms of magic and performance and taking a relational and narrative turn, *Ethics Inf. Technol.*, vol. 20, no. 2, pp. 71–85, 2018.

[81] D. J. Gunkel, A. Gerdes, and M. Coeckelbergh, Editorial: Should robots have standing? The moral and legal status of social robots, *Front. Robot. AI*, vol. 9, p. 946529, 2022.

[82] J. C. Gellers, *Rights for Robots: Artificial Intelligence, Animal and Environmental Law*. London, UK: Routledge, 2020.

[83] J. C. Gellers and D. J. Gunkel, Artificial intelligence and

international human rights law: Implications for humans and technology in the 21st century and beyond, in *Handbook on the Politics and Governance of Big Data and Artificial Intelligence*, A. Zwitter and O. Gstrein, eds. Cheltenham, UK: Edward Elgar Publishing, 2023, pp. 430–455.

[84] A. Birhane, Algorithmic injustice: A relational ethics approach, *Patterns*, vol. 2, no. 2, p. 100205, 2021.

Daniel Estrada is a university lecturer at the Department of Humanities and Social Sciences, New Jersey Institute of Technology, USA. He teaches courses in engineering ethics, AI ethics, and the philosophy of mind. His research interests lie at the intersection of agency and autonomy, computational complexity, and robot rights.

[85] A. Birhane, W. Isaac, V. Prabhakaran, M. Diaz, M. C. Elish, I. Gabriel, and S. Mohamed, Power to the people? Opportunities and challenges for participatory AI, in *Proc. Equity and Access in Algorithms, Mechanisms, and Optimization*, Arlington, VA, USA, 2022, pp. 1–8.

[86] I. Kavdir and D. E. Guyer, Apple grading using fuzzy logic, *Turkish Journal of Agriculture and Forestry*, vol. 27, no. 6, pp. 375–382, 2003.