PhiMiSci
Philosophy and the Mind Sciences | Vol. 6 | 2025

Attention and collective interests in artificial intelligence

In search of a regulatory framework

Carolyn Dicey Jennings[1,*]
Carlos Montemayor[2]

[1] University of California, Merced [2] San Francisco State University
* Primary contact: cjennings3@ucmerced.edu

Jennings, C. D., & Montemayor, C. (2025). Attention and collective interests in artificial intelligence: In search of a regulatory framework. *Philosophy and the Mind Sciences*, *6*. https://doi.org/10.33735/phimisci.2025.12164

Abstract

Recent debates frequently refer to artificially intelligent systems as agents, sometimes referencing their capacity for attention. Yet, the self-determination associated with agency requires a form of attention that is not yet present in artificial systems. It is thus worth asking how these artificial systems achieve the results they do. In this paper we explore the role of attention in artificial intelligence and argue that we should understand these systems as collective agents comprising the software developers, creators of training data, and users. Given this new understanding of artificial intelligence, we suggest a novel method for regulating these technologies: a corporate law approach. By treating artificially intelligent systems as corporate subcontractors, we can better protect both users and the broader public.

Keywords: Agency • Artificial intelligence • Attention • Collective action • Collective attention • Corporate law • Interests • Value alignment

1 Introduction

Attention is a hot topic in the search for artificial mental functions, such as artificial intelligence and artificial consciousness. It is, after all, implicated in nearly all mental functions, since it is the way that we manage and organize our mental resources. It is difficult to imagine how we could respond intelligently to a situation without first prioritizing certain aspects of that situation. Similarly, it is difficult to imagine being conscious of a scene without that scene being organized around an object or direction of focus. We support the general consensus that the development of general artificial intelligence will depend on the development of artificial attention.

Yet, departing from other accounts, we note the entanglement and mutual support of attention with our interests, where interests are propensities to seek out particular stimuli.[1] Interests are largely associated with biological organisms, which are assumed to have needs driving those interests. Such needs are absent in artificial beings, but another option is available to support artificial attention. While research on AI largely treats it as having stand-alone intelligence, we suggest treating AI as a collective involving its software developers, the creators of its training data, and its users. Instead of having its own interests and attention, we argue that AI depends on collective interests that drive a collective form of attention. We then provide some suggestions for how to regulate AI within this conceptual framework, arguing that we should treat AI systems as corporate subcontractors, with special protections for users and the broader public.

[1] We can separate these internal, driving interests from those that are merely external and descriptive, such as "ideal-regarding" interests (e.g. Feinberg, 1977) or "objective interests" (e.g. Railton, 1986), distinguishing this paper from the literature on interests within ethics; the latter types of interests would only be interests on our account if they were motivating factors for the individual in question. Another difference: much of the ethical discussion on interests concerns the contents or objects of interests and how to weigh them (e.g. Scanlon, 2000, Parfit, 2006), while our concern is the function of interests for the beings who have them and how those beings manage their interests.

2 Background and overview

Debates on the possibility of artificial general intelligence and artificial consciousness often turn on whether biology is essential to the mind (see, e.g., Seth, 2025). One version of the debate has to do with the role of language. Some see intelligence and consciousness as rooted in language (see, e.g. Dennett, 1991), and language as machine achievable, paving the way for artificial intelligence and consciousness. Others see intelligence and consciousness as present in non-linguistic creatures (see, e.g. Godfrey-Smith, 2016), leaving open the possibility that some aspect of our biology is key and potentially blocking the path to artificial forms.[2]

Recent technological advances have reopened these debates, with many expressing concern about the possibility of artificial general intelligence and artificial consciousness (see, e.g., Bojić et al., 2024). These advances are based on earlier developments in what is called "the attention mechanism." While in 2013 Helgason argued "attention has not yet been recognized as a key cognitive process of AI systems," this has now changed. In their landmark paper, "Attention is all you need," Vaswani and colleagues argued that "scaled dot-product attention" could replace core functions of neural networks more efficiently and with greater interpretability (2017); most "state-of-the-art" language models now use this mechanism (Sohn et al., 2024).

While attention is now widely recognized as an important part of AI development, the attention mechanism is not a true or complete form of attention. Recent models are said to include "an attention process that filters information" (Garrido-Merchán et al., 2020), but attention is more than a mere filter, and requires both bottom-up and top-down influences. That is, attention balances available stimuli and their relative salience (bottom-up influences) with the goals of the attending subject (top-down influences) in a flexible manner in order to best distribute mental resources. Shi, Darrell, & Wang (2023) argue that while "transformers still lack the ability of task-guided top-down attention," it is possible to develop a mechanism that

[2] Of course, we might yet *simulate* intelligence and consciousness without concern about artificial suffering or other facets of experience.

combines top-down and bottom-up inputs (see also Mittal et al., 2020).[3]

In keeping with this earlier work, we argue in this paper that artificial general intelligence will depend on an artificial form of attention. However, while significant progress has been made with mechanisms that approximate attention, we argue that a complete, stand-alone artificial attention will depend on the development of artificial *interests*. While interests are typically thought to be limited to the biological realm, we explore the possibility of non-biological, collective interests subserving a collective form of attention in AI. To get there we start by exploring the nature of interests for living beings (section 3) and the role of attention in managing interests in such beings (section 4). We then consider whether intelligence and attention are possible without interests (section 5). We find that the absence of interests would render attention chaotic and intelligence inefficient, while also creating problems for value alignment, which are already well known to beset AI. We introduce the possibility that AI operates like a collective, rather than having stand-alone intelligence, attention, and interests (section 6), and then discuss the special problems such a collective would introduce for value alignment (e.g. opacity; section 7). Finally, we consider how corporate law could help us to approach those problems (section 8) before concluding the paper. According to our approach, we can regulate AI without full alignment by attributing legal responsibility to AI systems independently of the value alignment problem.

3 The nature of interests for living beings

In the mind sciences interests are both ubiquitous and underspecified, causing Allport to claim decades ago that "one of our greatest defects is our lack of a consistent or adequate theory of interest"—a problem that is yet unsolved (1946, 341). Most generally, interests are directed at something and, all else being equal, push toward that thing until satiated. They are thought to be what drives biological organisms, similar to forces in physics (Valenstein, 1968, 5), and are more general than both needs and desires. Using Dretske's famous example of magnetosomes, one might say that marine bacteria with an interest in avoiding oxygen use internal magnets to move away from the oxygen-rich surface of the ocean (Dretske, 1986). More complex beings, such as humans, may have contrary interests, going toward the surface of the ocean when driven by an interest in breathing, for example, and away from it when driven by an interest in exploration.

Interests range from basic to advanced. They are associated with the survival of the organism but can also correspond with the survival of other organisms, particularly when they have shared genetic history. They are often used to confer moral status but can be discussed independently of moral questions (Holm, 2012). The vast majority of living beings are complex, with multiple interests; as mentioned above, these interests can be more or less unified. Attention is what allows minded creatures to manage their interests—to prioritize some interests over others (Jennings, 2022; this is discussed at length in the next section). We might thus assume that the capacity to constrain some interests for the sake of the whole collection of interests indicates the presence of attention, which can occur across the phylogenetic tree.

Individual interests come up both in discussions about the mind and about biology.[4] In discussions about the mind, interests are central to understanding the nature of attention, motivation, and learning (see, e.g., Berlyne, 1949; Hidi, 2006). In biology, interests are connected to the concept of life itself: "biological interests are interests that living things have merely in virtue of being alive" (McShane, 2021, 3499). Varner (1990) treats these as separable: "I argue that the best account of individual welfare would be one that recognizes the existence of *two* kinds of interests, *preference* interests on the one hand and *biological* interests on the other" (265). Yet, if

[3] Some have sought to do this through the "attention schema theory," using an attention schema, or a model of attention, to control the balance of top-down and bottom-up inputs. This is argued to both improve performance and to bring about a form of consciousness (Graziano and Webb, 2015). It remains unclear whether an explicit model of attention is necessary, or whether simpler mechanisms would suffice for both aims (Piefke et al., 2024).

[4] We are omitting for now social and political interests. We assume these to be analyzable as collective interests, which we discuss later in the paper.

one treats the mind as rooted in biology then one might expect continuity between the two—this is the approach we take in this paper.

How could this work? Decades ago a neuroscience working group sought a continuous account of *drives*, including an attempt to connect the more biological general drive (i.e. arousal) with more mental specific drives (e.g. fear). They found one region of the brain (the reticular activating system in the brain stem) to be responsible for general drive and another (the hypothalamus) to be responsible for specific drives (Valenstein, 1968). Specificity could come from interaction with a specific environment or from within the organism itself. Recently scientists have explored the possibility that genetic mutations (which lead to cell diversity, or "somatic mosaicism") can subserve the diversity of mental functions: "the complexity and unique features of the brain suggest that somatic mosaicism can play an important role in behavior and cognition" (Paquola et al., 2017; see also Muotri and Gage, 2006).[5] In our own work we treat the interests managed by attention as continuous with those of biology, both evolved and "evolving" (Montemayor and Haladjian, 2015; Jennings, 2022).

4 The function of attention for living beings with interests

The function of attention for living beings with interests is in allowing for their self-determination, which is especially important for complex beings with multiple interests. A common theme throughout work on interests is the claim that "all and only living organisms have interests" or "only individual organisms can plausibly be said to have interests" (Varner, 1990, 253). This idea is driven by the view that only organisms have self-determination: "the good of an artifact is really the good of something else, or at least derived from the good of something else; it is not a good that the artifact has in its own right" (McShane, 2021). The ability to determine one's own goods or interests is clearly absent from most artifacts. The question of this paper is, of course, whether AI is like other artifacts, which requires us to explore the reasoning behind the claim that organisms are unique in this regard. Do organisms determine their own interests in a way that is distinct from non-living things? We claim that this would require attention, at least for minded creatures like us.

A commonly accepted feature of all living beings is that they self-*organize* (Boden, 2000, 117). Yet, self-organization is also said to occur in non-living things, such as the sand patterns in deserts and shorelines (Coco and Murray, 2007), depending on the interpretation. Some see it through the lens of self-*production* or "autopoiesis," which is sometimes said to occur in non-living things (Boden, 2000, 123).[6] Others see it through the lens of self-*maintenance*, which is, likewise, not exclusive to living things (see, e.g., Holm, 2012).[7] But if self-organization occurs in non-living things it won't be able to support the claim that self-determination is exclusive to living beings. What's more, self-*determination* seems like it would require something closer to *transformation* or *control,* rather than mere production or maintenance. Thus, these accounts seem unlikely to show us why living beings are thought to be unique in this regard.

While the above accounts of self-organization include the concept of a boundary, missing is the concept of energy use across the boundary, which we take to be one of the critical elements of true self-determination. As Collier (2004) argues, any form of "self"-organization assumes individuation of the self-organizing system—a way of demarcating the system that self-organizes from that which surrounds the system. Any such demarcation requires internal cohesion relative to its environment. This internal cohesion, in turn, requires an "entropy gradient" across the boundary between the self-organizing system and its environment. That is, whereas

[5] Mutations are particularly preserved in the brain since most neurons last a lifetime (or longer; Magrassi et al., 2013),

[6] In living things, it is said to have three connected criteria: "(i) a boundary, containing (ii) a molecular reaction network, that (iii) produces and regenerates itself and the boundary" (Thompson, 2004, 386).

[7] "Self-maintenance is characterised as a property of systems that are able to exert a causal influence on their surroundings in order to maintain (at least some of) the boundary conditions required for their own existence."

disorder or entropy generally increases, internal cohesion requires that entropy is decreased for the ordered system. For that order to be maintained over time, energy is required. One could contrast this with a system in which energy is conserved: "the particles would just bounce right back to their original potential energies, and there would be no formation of organization" (Collier, 2004, 152). Even sand patterns—classic examples of self-organizing systems, mentioned above—are explained through "the continuous input of energy in the system (through waves, wind, and tides)" (Coco and Murray, 2007).

In living beings the use of energy across a boundary is explained through *metabolism*, or "the use, and budgeting, of energy for bodily construction and maintenance (and behaviour)" (Boden, 2000, 121). The use of energy by a living being is for the sake of that being's interests; we might thus see interests and metabolism as two sides of the same coin. Without metabolism across a boundary there is nothing "internal" driving the organism (no internal cohesion), and so no interests, and without interests there is no need for metabolism (no need to maintain internal cohesion): "metabolism denotes energy dependency, as a condition for the existence and persistence of the living thing as that particular physical unity" (Boden, 2000, 119). In Boden's view, artificial beings do not yet have the capacity for true metabolism, and so do not have the capacity for true self-organization. That is, the energy used by artificial beings is not used for the sake of self-construction and self-maintenance, as required by Boden to count as metabolism (Boden, 2000, 121).[8]

For those interested, the metaphysics of this can be seen through the concept of contextual emergence. "Emergence" is generally used to refer to phenomena that are not reducible to component parts and their local interactions. Contextual emergence is specifically said to occur when the parts or components of an emergent phenomenon are necessary but not sufficient for that phenomenon because a specific context is also necessary (Jennings, 2022; Bishop and Atmanspacher, 2006; see also Bishop et al., 2022). Contextual emergence is distinct from supervenience: supervenience allows for powers or properties that are shared across substrates, as the components are sufficient but not necessary for the supervening properties. In contrast, contextual emergence allows for powers that are unique to an individual, since the components are necessary but also dependent on a particular context to yield emergence. Contextual emergence avoids problems known to befall weak emergence, strong emergence, and supervenience: it is neither subject to the reduction problems of weak emergence (in which the components are necessary and sufficient for the emergent phenomenon) nor the interaction problems of strong emergence (in which the components are neither necessary nor sufficient). The reliance on context prevents the powers of the emergent phenomenon from being epiphenomenal, overdetermined, or circular, as in supervenience: the powers of the emergent phenomenon depend on both its components and the surrounding context, without that context being thereby subsumed by the emergent phenomenon as one of its components.

While metabolism allows for contextually emergent interests that drive the organism, *determination* of these interests by the organism requires something more. What allows an organism to determine their own interests is attention (Jennings, 2020, 2022). Attention describes the process of prioritizing preferred stimuli and responses over non-preferred stimuli and responses, potentially resulting in the selection of the preferred stimulus or response. This process allows organisms to prioritize some interests over others, as when they prioritize short-term interests (e.g. a current goal, such as exploration) over long-term interests (e.g. those relevant to survival, such as breathing), or vice versa. Attention thus allows organisms to shift and transform their own boundaries—to determine what will become part of the organism and what will not. This occurs in the short term because the prioritization of a particular interest helps determine what the organism takes in or consumes (e.g. preferred stimuli), and in the long term because interests that are not prioritized lose resources and power, resulting in less

[8] "Metabolism of this third type inevitably involves closely interlocking biochemical processes, to 'engineer' the organism's self-maintenance, growth, and activity. Because of the unavoidable tendency to disorder (i.e. the second law of thermodynamics), metabolism must involve continual energy intake from the outside world. And that, in turn, will engender further chemical processes within the body" (Boden, 2000, 121).

influence on the organism. The organism is essentially rewarding one part of itself through attention, and that reward has downstream impacts that shape the organism. Hence, through attention the organism takes part in self-determination.

To see how this might work for humans we can use the concept of contextual emergence at another level. That is, the management of an organism's interests can be understood as an emergent phenomenon concerning the full collection of interests. In the human case interests at least partly correspond with brain activity and the relevant context is the thermodynamic disequilibrium enabled by the brain, brain stem, and skull: the brain consumes energy through the brain stem and this process is protected by the skull. We can thus see attention as emerging in a specific context to enable the management of our interests. This disequilibrium is analogous to that allowed by metabolism, which makes way for interests or drives in the organism. These interests or drives are "internal," coming from the organism rather than the environment. The disequilibrium of the brain, brain stem, and skull allow this internalization to occur at another level, with organizational powers *over* the organism, and so internal *to* the organism. Thus, the kind of energy consumption that supports life through metabolism can also support the mind through attention.

5 The importance of interests for attention and intelligent behavior

While attention is what manages interests in living beings like us, one might argue that attention could occur without interests, and so without metabolism. Perhaps attention could be described as the management of information, rather than the management of interests. This would allow us to extend the concept of attention to artificial beings that do not have interests, but do have information.

There are two problems with this. The first is conceptual. There are multiple senses of "information." While attention is often discussed in terms of information, this information is assumed to have material grounding and resource implications. As Collier (2004) puts it in his argument on the energy requirement for individuation, "these principles are not restricted to material and energy flows, but apply to information flows as well, as long as the information is understood in terms of its material basis" (154). Another sense of the term is "Shannon information," which is understood as mere variation. Variation is of course possible without the energy exchange enabled by metabolism. However, selection cannot be based on mere variation, but only variation of a particular type (e.g. a particular spatial scale). Thus, if artificial beings do not have their own interests—some standard by which to judge this variation—their "attention" will depend on the interests of others, such as the interests of their developers and the owners of the data used to train them.

The second is pragmatic: to be useful artificial beings need to have a certain form of direction (e.g. to solve the "frame problem").[9] While researchers have begun to solve this pragmatic issue through selective and filtering mechanisms that approximate attention, the lack of internal management and coordination resists internal alignment. This and the above problem both cause downstream issues with value alignment that are by now infamous, as when unwanted interests embedded in data influence AI outputs or when AI acts in surprising ways against the wishes of its developers (see, e.g., Montemayor, 2023).

Intelligence is likewise bound by pragmatism. Intelligence is ubiquitous in nature, often defined under the general characterization of optimal problem solving. But this characterization is too abstract to capture the unique source of intelligence in biological organisms. Problem solving concerns a very large set of possible problems, only a few of which are worth solving, and even of this much smaller set, the select problems that must be solved must be addressed in very specific ways. An intelligent agent doesn't need massive collections of data and enormous energy resources because this selection of processes occurs quite efficiently. The same holds for how intelligent agents communicate. A truly intelligent agent never

[9] The Frame Problem has multiple versions, but we mean the most general sense of the problem, which is the problem of how to limit the scope of possible computations given limited processing capacity (see, e.g., Shanahan, 2016).

just optimizes solutions to randomly selected communication problems, or needs to be prompted to be truthful and commonsensical.

Genuine intelligence fundamentally includes *problem selection* according to a hierarchy of interests that only really intelligent organisms have. The problems an intelligent agent solves are determined by her interests—she autonomously determines which interests are at the top of the priority scale. No one has to select them for her or prompt her to solve them in a specific way. This version of the frame problem is not merely functional (i.e., if I want to buy this ticket, which forms do I have to fill out, or what is the best strategy for getting a good seat at the event). Rather, it involves a more complex prioritization of goals that depends on the interests of the agent, but which need not be explicit or conscious. This is where being alive and having a boundary with respect to the environment matters. A substantial number of the interests a living agent has are determined by the kind of body, or metabolic boundary, that she has, including the complexity of environmental and representational needs she must fulfill to maintain her life and cognitive well-being. In humans, interests are deeply shaped by social pressures and commitments. Factors that shape the interests of a living agent include her lifespan, social relations, social expectations, and demands on her learning capacities.

Thus, intelligence requires interests to determine which problems it should tackle, and in what way. While these interests can be coded or built into the architecture, this prevents the balance of stability and flexibility that is characteristic of intelligent beings. Instead of a hierarchy of interests that is driven by an agent with needs in a particular environment, artificial beings are either overly rigid, determined by preset "goals" or tasks (but without any corresponding "drive"), or overly chaotic, with no principled way to determine which goal or task should take priority. True intelligence comes from an agent having its own interests and goals, enabling curiosity and meta-learning (learning what and how to learn), with resource availability naturally directing the agent toward either stability or flexibility. This demonstrates the connection between intelligence and attention: because problem selection depends on interests, the attention capacities of the agent are essentially involved in how she intelligently solves problems. Attention, as goal-prioritization and selection, is what distinguishes functional aspects of agency from full agency, thereby making a difference in how the selection of goals is fundamentally dependent on the agent.

To see this, consider recent advancements in "agentic AI." Agentic AI is growing in popularity, and most companies developing AI have different models of AI "agents" with various degrees of autonomy (Kasirzadeh and Gabriel, 2025). While agentic AI promises to take on many managerial tasks autonomously, this kind of autonomy is a form of "functional" autonomy: "In the context of AI agents, agency is best understood in terms of the capacity to perform actions *without external direction or control,*" which "does not require that agents have mental states that are analogous to those of human beings" (Kasirzadeh and Gabriel, 2025, 3–6). This form of "autonomy" lacks the characteristic integration and control of human and animal autonomy. Humans decide what are the most optimal steps to take in order to fulfill a task, but only once they have prioritized a goal—is this goal urgent, is it important, and to what extent? Similarly, LLMs have capacities that resemble linguistic skills for maintaining a conversation, identifying speech acts, and reliably producing context-sensitive answers, but seem to be incapable of satisfying the requirements for full conversational alignment (Sterken and Kirkpatrick, 2025). Ultimately, this is an issue of who is in control of what gets prioritized as a goal that is worth pursuing. Goal-prioritization is a fundamental aspect of human agency, and no amount of goal application can replace it. It is the choice we have over goal-prioritization that makes us responsible agents in all domains (e.g., legal, ethical, practical, and epistemic), and which makes our lives free and worth living. This kind of autonomy is not present in agentic AI.[10]

[10] If it were ever developed, it would likely constitute a serious risk to our own autonomy (Russell, 2019); we are not suggesting that here.

6 Introducing collective interests and collective attention

As mentioned in the section above, the lack of attention and interests creates problems for value alignment, or problems with bringing AI into alignment with our system of values, whether aesthetic, epistemic, moral, political, or legal. As Bojić, Stojković, and Jolić Marjanović (2024) put it: "The rapid advancement toward superintelligence requires continuous monitoring of AI's human-like capabilities, particularly in general-purpose models, to ensure safety and alignment with human values." Yet, it isn't clear how to achieve this alignment if AI has neither its own interests nor its own attention. In humans, it is attention that allows for internal alignment, while having similar interests allows us to be aligned in our values.[11] We share many of our interests with animals, since they also have interests that depend on their bodily boundary and metabolic needs. While the prevailing characterization of the alignment problem in AI focuses on ethical value, the alignments of intelligent beings include various kinds of other alignments on which our well being depends. For instance, our interests in communicating efficiently, representing the environment jointly with others in order to interact and cooperate with them, and so on. Our alignment in interests is the basis for cooperation and social intelligence.

Here attention is important as well. Without reliable attentive capacities the agent will fail at identifying salient problems that must be solved in the context of social cooperation. An important example is that of "cost-avoidance," in which a pressing interest commands our attention because the cost of not attending to it is high. Consider leaving on time for a meeting. Whether we attend to this issue depends on the timing of other events and the relative ranking of our interests. Very high on our ranking is our own life maintenance (e.g. eating a snack before we leave) but for generally intelligent agents like us, mere life maintenance is not sufficient. Our well being includes a large variety of interests that can be placed at the top of the hierarchy of what we value and need to achieve, depending on our circumstances. Without a hierarchy of interests AI is unable to make these sorts of judgments, and to engage in cooperative endeavors that require give and take.

As we have tried to demonstrate, since contemporary AI has neither a body boundary nor autonomous interests of any kind, it lacks genuine intelligence. For AI to be intelligent, it would need to have interests and attention capacities. Recently, Man, Damasio and Neven (2022) argued that *needs* are crucial for intelligence, particularly needs that concern the vulnerability of living beings to their environment. We agree with them about the importance of the body boundary and its relation to interests, but they make this point mainly in terms of vulnerability. We think that it is not merely our bodily vulnerability that makes us intelligent—rather, our interests, organized by attention, make us intelligent.

This problem of value alignment shows how critical it is that we understand the role of interests in AI, even if they are not self-determined or guided by attention. AI is making an impact, but who or what is driving that impact? There are three possibilities here. First, AI could have its own, entirely independent interests. We think this possibility is ruled out by its lack of metabolism or equivalent, which is a problem that is unlikely to be solved anytime soon. Boden (2000) argues this point for artificial life: "strong [artificial life] is indeed impossible, because of the lack of (third-sense) metabolism and a self-constituted bodily boundary" (142). Again, while Boden allows that artificial systems use energy, they do not do so for the sake of self-organization, the third sense of metabolism: "Metabolism (in the third sense) is the use of energy-budgeting for autonomous bodily construction and self-maintenance" (122). As argued above, without energy use across a boundary we do not get interests that belong to that particular entity, interests that are self-determined and attributable to the agent as such.

Second, AI could be driven by the interests of its users, which is arguably how other digital technologies help us to solve problems (e.g. calculators). This is a good starting point, but an insufficient explanation of how AI actually works, which is unlike other forms of digital technology.

[11] We are setting aside for the moment the possibility of collective attention, which we later consider as another route to alignment in social groups.

While other forms of digital technology arguably serve as extensions of mental storage (i.e. memory) or skill (e.g. navigation), AI extends thinking itself, generating its own information and solutions. Consider the difference between "Googling" a question in 2023 and putting that question to AI: in the first case the response will be a list of static resources provided by others with their own interests, whereas in the second the resource will be *created* in virtue of *the question itself*. The dynamic nature of AI prevents us from treating it as driven entirely by us.

Thus, in the case of AI we believe we are facing the third possibility: the interests of AI are based in a collective, an idea initially suggested by Norbert Wiener, an early AI pioneer (see, e.g., Montemayor, 2023, 170–174). Specifically, the user's interests are working in concert with interests embedded in the AI system. These are not autonomous interests that belong to the AI, but echoes of the interests of its developers and the creators of its training data. Compare this to how our own interests work. A momentary interest in cycling might lead us to prefer an image of a bicycle to one of a car, but a longstanding interest can lead to us perceiving bicycles more readily, speeding our automatic reactions to them. This is the difference between "fluid" interests that are maintained by the executive and "formed" interests that depend on long term memory systems (Jennings, 2022). But we also have "fixed" interests that are set in our physiology, inherited from others (e.g. an interest in "redness"; Jennings, 2022). Fixed and formed interests can direct our behavior without autonomous, fluid interests, but they are only *ours* in virtue of integration with our fluid interests. Similarly, AI expresses the interests of its developers and the creators of its training data without having its own interests, and when users work in concert with AI they are essentially working *with* or *through* the interests of these other people, operating as a collective. Thus, we might understand this as a collective form of attention.

Compare this to a more familiar human collective—a club, such as a book club. When a small group of people gets together to discuss a book, we might see the time spent on different topics as due to their collective attention. That is, we can see time as a resource that is allocated to topics of the highest priority, and away from topics of lowest priority. Each individual has their own interests, but the collective interests are those held in common across individuals, which dictate the flow of conversation. This is true even if there is a book club "host": the host can try to direct the conversation to topics of personal interest, but the group will inevitably exert its own pressures. Similarly, the output of AI is determined by the combined interests of the user, the AI developer, and the creators of its training data (i.e. both the owners of that data and those who compile the data)—a massive collective. The process of rendering an output as well as the structure and contents of that output rely on interests beyond those of the user.

Left unclear is how we should understand this new collective of users, data creators, and developers: are their interests merely aggregated to solve problems or are there new, emergent interests that belong to the AI collective? That is, does the AI collective have interests that go beyond those of the users, creators, and developers? This need not involve strong or even contextual emergence: supervenience would come with new properties at the level of the collective that may not be predictable from its components. While this possibility may sound fanciful, corporations are often conceived as "ruthlessly pursuing their own goals and disregarding the interests of their members" (List and Pettit, 2011, 129), a possibility that is enabled by their intrinsic structures (see also Montemayor, 2023, 171–174). Such interests might align with ours, but may also differ substantially from our own, given that they are emergent. These emergent interests may be particularly difficult to apprehend and assess, since they include the interests that emerge from compressing information from data that belong to all of us, and which need not align with the more narrow interests of shareholders and developers at AI companies.

Take a simplified example: imagine the use of AI by a college student with an interest in writing a paper in order to pass a class. In this case, let's presume that the data used to train the AI were thousands of other papers by college students, and that the data reflect a general interest in obtaining a passing grade for minimal work output (e.g. vague language, multiple errors). Let's presume that the AI developers likewise want to sell their product to students on the grounds of that interest: to write a passing

grade for minimal effort. Errors in the data introduced by students trying to minimize effort may be ironed out by the large sample of papers. Yet, past papers cannot predict future topics, so developers will need to encourage some extrapolation while also minimizing the introduction of new errors that could bring the user grade below passing. Thus, the combination of these interests might lead to a new collective interest of "sufficient hallucination": just enough "hallucination" to allow the AI to be used on new questions or topics without substantially reducing the basis of truth in the papers that allowed them to obtain a passing grade (see, e.g. Kalai et al., 2025). Note that none of the individuals involved have an interest in sufficient hallucination—this comes out of the tension between the various interests involved.

So, in this third case, AI operates as a collective that either has none of its own interests (borrowing entirely from the interests of developers, data creators, and users) or has its own interests that emerge from the collective. In either case value alignment will require reference to the full collective, rather than either the user or the AI software on its own. We consider below how to proceed with this problem following some initial considerations on the unique nature of the AI collective.

7 The problems of opacity and projection

We think recognizing AI as a collective—with collective interests driving a collective form of attention to allocate resources—best fits how it currently operates. Yet, we also think this recognition is insufficient for value alignment. In order to align values one must have both (implicit or explicit) knowledge about those values and some sort of control over them. The fact that AI operates as a collective raises special problems with both knowledge and control that are not seen in other collectives.

First, there is a problem of opacity. When we operate as a collective we typically have some access to others in the collective that would provide us with knowledge about their values and interests. This is clear in the book club example, but even a group of strangers in the same physical location will know something about the values of other strangers at that location, since those values led them to that location. Strangers on a train, for example, will know that other strangers on the train value train travel. With AI this is not the case. Users of AI typically know little to nothing about the values of either the software developers behind the AI or those of the creators of the data used to train the AI beyond their status as other humans. Digital information already transcends physical location and AI cuts the link to the origin of that information. If users of AI do not have epistemic access to the values of other members of the AI collective, they cannot expect value alignment with that collective (see also Arvan, 2024).

Second, there is a problem of value projection. In the face of opacity we sometimes project an explanation of what is happening based on our own values (see also Vallor, 2024). We might, for example, project emotional warmth onto AI when it is not capable of such warmth. The consequences of that go beyond mere epistemic ones, including unhealthy emotional attachment; multiple papers are beginning to report addictive behaviors with AI, especially for the heaviest users (Fang et al., 2025). There are also risks concerning empathic AI, in which the illusion of empathy puts users in a situation of potential abuse and immediate deceit (Montemayor, 2023). Thus, this projection deepens the problem of opacity while also making the user less capable of controlling the values of AI in order to bring about value alignment.

These problems do not assume an emergent collective, but such emergence can increase opacity, if not projection. Consider, for example, what it is like to work in a large institution, such as a corporation or a university. Due to emergence[12], many of the ways the institution operates can be opaque, without a single person that could explain its full functionality. Similarly, if AI is an emergent collective then we should not expect anyone—its users, creators, or the creators of its training data—to understand all of its interests and values.

One might object that AI is no more plagued by these problems than our own minds. The unconscious mind is thought to be directed by interests and

[12] We are including supervenience as a type of emergence here.

values unknown to us, and we often rationalize our behavior by projecting consciously held values. We might see AI as no more a "black box" than our own unconscious, with interests and values inherited and set in our physiology without our personal oversight and control. This analogy fails because we do have some degree of control over our unconscious interests and values and because adult humans have had decades to adjust the impact of these interests and values. In contrast, the influence of a user's interests on contemporary AI is partial in the best case. We might compare it to remotely operating the brain of an alien, rather than our own brain—an alien with unknown background and agenda, marketed to seem safely aligned with our interests and values.

These considerations lead to a puzzle with respect to AI policy. Namely, if this is the right way to think about AI, how might we regulate it? After all, it is unlike both other forms of software and other forms of collectives. We argue below that we can get purchase on this by treating AI like a corporate entity that includes its developers, the owners of its training data, and its users. More precisely, AI systems would be corporate entities with a dependency relation to both the corporation that produced them and their users. This is compatible with opacity and rapidly changing situations, in which "residual control rights" could be part of the solution to the problem of AI regulation.[13] AI systems would then be corporate subcontractors for the purpose of legal responsibility, rather than full agents that require alignment with human values. This way, we can take on the legal challenge of regulating AI directly, without needing to address the thorny issue of alignment—including alignment with the interests of shareholders.

8 Corporate law as a solution

As mentioned above, Wiener suggested a link between AI and corporate entities as far back as 1950: "He anticipated the dangers of creating artificial superintelligences with goals not necessarily aligned with our own. What is now clear, whether or not it was apparent to Wiener, is that these organizational superintelligences are not just made of humans, they are hybrids of humans and the information technologies that allow them to coordinate" (Hillis, 2019). Wiener's argument about AI being equivalent to a corporate entity allows us to articulate forms of alignment that do not depend on specific interests and stances (Gabriel and Keeling, 2025), the mental alignment of preferences and goals (Arvan, 2024), or different scales of urgency (Kasirzadeh and Gabriel, 2025). The current framework for how corporations should protect customers, society, and human rights is available to us. We can start regulating AI systems as the largest corporate entities in human history.

AI systems are the largest corporate entities ever created due to both the large number of users and the large number of producers of data that are needed to fuel machine learning. AI companies extract this data without authorization from the producers of that data, the users, and the public in general to compete against other companies, leading to significant accumulation of capital. This aspect of corporate consumption requires urgent regulation, which has already begun in the case of copyright infringement (see, e.g., Lucchi, 2024). Beyond copyright, because users are part of the data generation and training of LLMs, we could compare them to workers, the way Uber drivers become part of the larger company even though they are not treated as full employees (see, e.g., Dubal, 2020). At 400 million current users, ChatGPT vastly outnumbers the employees, official and independent, of any company, which justifies regulating AI systems as comprising the largest corporate entities on the planet (Curry, 2025).

Since alignment with contemporary AI cannot occur through the normal channels of shared attention and interests, we must look for other possible sources of regulatory principles that could help us control and safeguard our interactions with AI. Fortunately, we have a very sophisticated set of principles to align collectives like AI systems available through existing corporate and civil legislation. By treating AI systems like corporate entities instead of individual agents we can better understand their interests, which are based in a collective, and better align them with civil

[13] "Actual contracts […] are poorly worded, ambiguous, and leave out important things […] who has the right to decide about the missing things? We called this right the residual control or decision right" (Hart, 2017, 1732).

and administrative law. However, AI technology presents us with unique, and well-known challenges, such as opacity, or a lack of clarity concerning how they arrive at conclusions, and systematic biases and "hallucinations." For this reason, one must also regulate the *interfaces* between AI and its users. These regulations can be less rigid and more programmatic, following the distinction between hard and soft law.[14] Soft law for interfaces can implement principles and recommendations regarding shared expectations that go beyond corporate and copyright law. These could be enforced by agencies that require either standard designs of interface applications, for instance to prevent exposure to addictive contents, or better ways of informing consumers. They could also involve recommendations for preventing various kinds of surveillance and privacy infringements that are the consequence of how informational interfaces are designed.

So far we have discussed two groups that would be better protected by this approach: the creators of training data, who are currently underacknowledged contributors to the AI enterprise, and the users, who are both partly responsible for the products they create with AI (e.g. plagiarized scholarship) and vulnerable to AI systems in their use (e.g. in the case of addictive content). A third group that deserves better protection is the broader public. For this, we appeal to enforceable legislation concerning corporate entities and their subcontractors. Importantly, existing precedents are insufficient. Precedents like Waymo v. Uber-Ottomotto concern existing law governing liability and copyright. These are standard cases of law enforcement between two rival companies that disagree about a fundamental transaction governed under the law. However, enforcement is a big problem for current efforts to regulate AI more generally, beyond the disagreement of companies, as the law cannot be easily applied to copyright infringements that are involved in cutting edge techniques that are also protected as industrial secrets. This is why a different type of legal mechanism is required for treating AI systems as corporate affiliates—a kind of corporate subsidiary that is a de-facto representative of the company without having individual agency. With respect to liability, similarly to the food industry, even if the public ignores the intentions or interests of AI affiliates (just as they ignore the inner workings of food companies), if they cause harm they can be held responsible for harming public interests.

We believe that this approach will make AI systems (and the companies that produce them) accountable, while dispelling the unproductive metaphor that they are agents because they are conscious persons or genuinely agentic individual subjects that deserve rights. The responsibility of AI systems is collective in the sense that their responsibility does not necessitate the standard capacities of individual human beings or any kind of alignment between an AI system and an individual human being, let alone the collective alignments of shareholders or of humanity as a whole. Instead of looking for abilities or individual agency, the responsibility of AI systems should be treated as corporate in the sense that companies are responsible for the actions of the AI systems they create, but in different ways and in different contexts, similarly to a corporate sub-contractor.[15]

Thus, without suggesting a solution to the alignment problem, we propose an initial implementable proposal to better align AI systems with current legislation and the principles of jurisprudence. OpenAI's dissolution of its original board and ethics team demonstrates the urgency of arriving at a different approach. Despite changing from a non-profit to a competitive and wealthy company, the nature of the risks the AI systems produced by OpenAI are independent of shareholder's interests, and this explains why they could be conceived as affiliates or subcontractors that present their own kinds of liability and harm. AI systems execute actions that are under the initial control of AI companies. Although their actions are independent and reflect different collective interests, including those of the user, they are acting on behalf of their master company, and this makes them de facto

[14] For many agreements and legally based recommendations that are not legally binding, the term "soft law" is used to indicate that, despite not being enforceable, such recommendations reflect the considered opinion of legal experts. "Hard law" concerns legislation and established jurisprudence that is binding and enforceable. This distinction is particularly useful in international law, which should also govern AI development given the global impact it is having on humanity.

[15] We thank an anonymous reviewer for this suggestion. Of course, since the user is also directing the AI system they are, too, partially responsible for its outputs or products.

representatives of the company and quasi-contractors performing a task that has been given to them.

Other corporate structures are possible, in which more explicit emergent interests that align with shareholders may be available. But as long as AI systems act with a degree of autonomy that is responsive to the tasks given to them by a company, they will act as their representatives and quasi-contractors of that company.

This legal framework could then provide a balance of stability and acceleration. AI companies need the stability of the slow systems of legal protections provided by civil and criminal law. Trust can only be created by the parts of the slow system that create the foundational conditions for equality and justice. Corporate law can move faster, and for the quickly disruptive effects of AI, a combination of corporate and soft law approaches could be our best option. The slow moving parts of our legal systems are getting disturbed by the acceleration of markets and technologies produced by AI companies. This puts us in a paradoxical legal situation, in which the slow systems of civil law that protect these companies are being challenged by the speed with which they are altering the public sphere and with it, public trust.

Technology is a source of rapid transformation, but also of accelerated disruption regarding systems of trust. No other corporate structure has accumulated the amount of wealth and size in terms of members and global impact than AI companies. For their success, they completely depend on the slow-moving legal system that protects their property. The law needs to react somehow to the disruption they are causing, since without their legal protections, grounded on public trust, they would fail. However, if they continue disrupting social systems, the potentially destructive consequences of their actions will not only affect them, they will affect all of us. This is why we should start regulating AI systems as corporate representatives of AI companies, subcontractors that have an unprecedented level of access and resources to potentially harm (or benefit) the public.[16]

[16] A full account of how to align values and interests between humans and AI systems goes beyond a corporate law approach, which is the main focus of this paper. We want to point out that such an effort requires many more actors: a combination of NGOs and non-for-profit organizations with heavy intervention and constant feedback from public groups. This is the only way to achieve real alignment and democratically achieved consensus.

9 Conclusion

We began this paper with questions about how AI is connected to attention, and ended by exploring how we should manage AI in the legal system. Our overall argument is simple. In our view, true intelligence requires attention, which requires interests. Thus, artificial intelligence will depend on interests of some kind, but does not currently have any of its own. We suggest thus treating artificial intelligence as dependent on the interests of its developers, the owners of its training data, and its users. This "collective agency" conception of AI leads to unique problems concerning control and regulation, but these can begin to be solved through corporate law, treating agentic AI systems as subcontractors. Thus, this paper is both conceptual and pragmatic: we aim to contribute to discussions of what it means to be an AI agent, as well as to contribute to discussions about what to do about AI.

AI is nimble, and so must be this paper: it does not cover all of the relevant literature, nor answer all of the relevant objections one might have. Our understanding of "interests" and other concepts may seem unusual to some readers, while our invocation of "contextual emergence" may surprise others. We do not seek to satisfy everyone, but we do hope that the ideas here can lead to clarity in debates on the role of AI in our social, political, and legal systems.

Acknowledgements

Thanks to our managing editor, Sascha Benjamin Fink, an advising editor, Michael Pohl, and two anonymous referees for helpful feedback that greatly improved the paper.

References

Allport, G. W. (1946). Effect: A secondary principle of learning. *Psychological Review*, *53*(6), 335–347. https://doi.org/10.1037/h0059295

Arvan, M. (2024). 'interpretability'and 'alignment'are fool's errands: A proof that controlling misaligned large language models is the best anyone can hope for. *AI & Society*, *40*, 3769–3784. https://doi.org/10.1007/s00146-024-02113-9

Berlyne, D. E. (1949). Interest as a psychological concept. *British Journal of Psychology*, *39*(4), 184–195. https://doi.org/10.1111/j.2044-8295.1949.tb00219.x

Bishop, R. C., & Atmanspacher, H. (2006). Contextual emergence in the description of properties. *Foundations of Physics*, *36*(12), 1753–1777. https://doi.org/10.1007/s10701-006-9082-8

Bishop, R. C., Silberstein, M., & Pexton, M. (2022). *Emergence in context: A treatise in twenty-first century natural philosophy*. Oxford University Press. https://doi.org/10.1093/oso/9780192849786.001.0001

Boden, M. A. (2000). Autopoiesis and life. *Cognitive Science Quarterly*, *1*(1), 115–143.

Bojić, L., Stojković, I., & Jolić Marjanović, Z. (2024). Signs of consciousness in AI: Can GPT-3 tell how smart it really is? *Humanities and Social Sciences Communications*, *11*, 1631. https://doi.org/10.1057/s41599-024-04154-3

Coco, G., & Murray, A. B. (2007). Patterns in the sand: From forcing templates to self-organization. *Geomorphology*, *91*(3-4), 271–290. https://doi.org/10.1016/j.geomorph.2007.04.023

Collier, J. (2004). Self-organization, individuation and identity. *Revue internationale de philosophie*, *228*(2), 151–172. https://www.jstor.org/stable/23955622

Curry, D. (2025, September). ChatGPT revenue and usage statistics. https://www.businessofapps.com/data/chatgpt-statistics/

Dennett, D. C. (1991). *Consciousness explained* [P. Weiner, Illustrator]. Little, Brown; Co.

Dretske, F. (1986). Misrepresentation. In R. J. Bogdan (Ed.), *Belief: Form, content, and function* (pp. 17–36). Oxford University Press.

Dubal, V. B. (2020). An uber ambivalence: Employee status, worker perspectives, and regulation in the gig economy. In D. Das Acevedo (Ed.), *Beyond the algorithm: Qualitative insights for gig work regulation* (pp. 33–56). Cambridge University Press.

Fang, C. M., Liu, A. R., Danry, V., Lee, E., Chan, S. W., Pataranutaporn, P., Maes, P., Phang, J., Lampe, M., Ahmad, L., & Agarwal, S. (2025). How AI and human behaviors shape psychosocial effects of chatbot use: A longitudinal randomized controlled study. https://doi.org/10.48550/arXiv.2503.17473

Feinberg, J. (1977). Harm and self-interest. In P. M. S. Hacker & J. Raz (Eds.), *Law, morality and society: Essays in honour of h.l.a hart* (pp. 285–308). Oxford University Press.

Gabriel, I., & Keeling, G. (2025). A matter of principle? AI alignment as the fair treatment of claims. *Philosophical Studies*, *182*(7), 1951–1973. https://doi.org/10.1007/s11098-025-02300-4

Garrido-Merchán, E. C., Molina, M., & Mendoza, F. M. (2020). An artificial consciousness model and its relations with philosophy of mind. https://doi.org/10.48550/arXiv.2011.14475

Godfrey-Smith, P. (2016). *Other minds: The octopus and the evolution of intelligent life*. William Collins.

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. *Frontiers in psychology*, *6*, 500. https://doi.org/10.3389/fpsyg.2015.00500

Hart, O. (2017). Incomplete contracts and control. *American Economic Review*, *107*(7), 1731–1752. https://doi.org/10.1257/aer.107.7.1731

Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, *1*(2), 69–82. https://doi.org/10.1016/j.edurev.2006.09.001

Hillis, D. (2019). The first machine intelligences. In J. Brockman (Ed.), *Possible minds: 25 ways of looking at AI* (pp. 170–177). Penguin Random House.

Holm, S. (2012). Biological interests, normative functions, and synthetic biology. *Philosophy & Technology*, *25*, 525–541. https://doi.org/10.1007/s13347-012-0075-6

Jennings, C. D. (2020). *The attending mind*. Cambridge University Press. https://doi.org/10.1017/9781108164238

Jennings, C. D. (2022). *Attention and mental control*. Cambridge University Press. https://doi.org/10.1017/9781108982269

Kalai, A. T., Nachum, O., Vempala, S. S., & Zhang, E. (2025). Why language models hallucinate. https://doi.org/10.48550/arXiv.2509.04664

Kasirzadeh, A., & Gabriel, I. (2025). Characterizing AI agents for alignment and governance. https://doi.org/10.48550/arXiv.2504.21848

List, C., & Pettit, P. (2011). *Group agency: The possibility, design, and status of corporate agents*. Oxford University Press.

Lucchi, N. (2024). ChatGPT: A case study on copyright challenges for generative artificial intelligence systems. *European Journal of Risk Regulation*, *15*(3), 602–624. https://doi.org/10.1017/err.2023.59

Magrassi, L., Leto, K., & Rossi, F. (2013). Lifespan of neurons is uncoupled from organismal lifespan. *Proceedings of the National Academy of Sciences*, *110*(11), 4374–4379. https://doi.org/10.1073/pnas.1217505110

Man, K., Damásio, A. S., & Neven, H. (2022). Need is all you need: Homeostatic neural networks adapt to concept shift. https://doi.org/10.48550/arXiv.2205.08645

McShane, K. (2021). Against etiological function accounts of interests. *Synthese*, *198*(4), 3499–3517. https://doi.org/10.1007/s11229-019-02293-8

Mittal, S., Lamb, A., Goyal, A., Voleti, V., Shanahan, M., Lajoie, G., Mozer, M., & Bengio, Y. (2020). Learning to combine top-down and bottom-up signals in recurrent neural networks with attention over modules. *Proceedings of Machine Learning Research*, *119*, 6972–6986. https://proceedings.mlr.press/v119/mittal20a.html

Montemayor, C. (2023). *The prospect of a humanitarian artificial intelligence: Agency and value alignment*. Bloomsbury Academic. https://doi.org/10.5040/9781350353275

Montemayor, C., & Haladjian, H. H. (2015). *Consciousness, attention, and conscious attention*. MIT Press. https://doi.org/10.7551/mitpress/9780262028974.001.0001

Muotri, A. R., & Gage, F. H. (2006). Generation of neuronal variability and complexity. *Nature*, *441*(7097), 1087–1093.

Paquola, A. C., Erwin, J. A., & Gage, F. H. (2017). Insights into the role of somatic mosaicism in the brain. *Current opinion in systems biology*, *1*, 90–94. https://doi.org/10.1016/j.coisb.2016.12.004

Parfit, D. (2006). Rights, interests and possible people. In H. Kuhse & P. Singer (Eds.), *Bioethics: An anthology* (2nd, pp. 108–112). Wiley-Blackwell.

Piefke, L., Doerig, A., Kietzmann, T., & Thorat, S. (2024). Computational characterization of the role of an attention schema in controlling visuospatial attention. https://doi.org/10.48550/arXiv.2402.01056

Railton, P. (1986). Moral realism. *The philosophical review*, *95*(2), 163–207. https://doi.org/10.2307/2185589

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Scanlon, T. M. (2000). *What we owe to each other*. Belknap Press.

Seth, A. K. (2025). Conscious artificial intelligence and biological naturalism. *Behavioral and Brain Sciences*, 1–42. https://doi.org/10.1017/S0140525X25000032

Shanahan, M. (2016). The frame problem. In E. Zalta (Ed.), *The stanford encyclopedia of philosophy* (Spring 2016 Edition). https://plato.stanford.edu/archives/spr2016/entries/frame-problem/

Shi, B., Darrell, T., & Wang, X. (2023). Top-down visual attention from analysis by synthesis. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2102–2112. https://doi.org/10.1109/CVPR52729.2023.00209

Sohn, G., Zhang, N., & Olukotun, K. (2024). Implementing and optimizing the scaled dot-product attention on streaming dataflow. https://doi.org/10.48550/arXiv.2404.16629

Sterken, R. K., & Kirkpatrick, J. R. (2025). Conversational alignment with artificial intelligence in context. *Philosophical Perspectives*, *38*(1), 89–102. https://doi.org/10.1111/phpe.12205

Thompson, E. (2004). Life and mind: From autopoiesis to neurophenomenology. a tribute to francisco varela. *Phenomenology and the cognitive Sciences*, *3*(4), 381–398. https://doi.org/10.1023/B:PHEN.0000048936.73339.dd

Valenstein, E. S. (1968). *Biology of drives - a report of an NRP work session: Concepts and experimental data on biological mechanisms of drives, motivation, reinforcement, and learning* (tech. rep. No. NASA-CR-103225). https://ntrs.nasa.gov/citations/19690022316

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press. https://doi.org/10.1093/oso/9780197759066.001.0001

Varner, G. E. (1990). Biological functions and biological interests. *The Southern Journal of Philosophy*, *28*(2), 251–270. https://doi.org/10.1111/j.2041-6962.1990.tb00545.x

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. https://doi.org/10.48550/arXiv.1706.03762