This is a forthcoming chapter in the *forthcoming* edited collection: Müller, Vincent C.; Dewey, Aliya R; Dung, Leonard and Löhr, Guido (eds.), Philosophy of Artificial Intelligence: The State of the Art (Synthese Library, Berlin: SpringerNature). If the reader wishes to cite it, please contact the author at giacomo.f.talamanca[at]gmail.com.

From AI to Octopi and Back.

AI Systems as Responsive and Contested Scaffolds

Giacomo Figà-Talamanca

**Abstract**

In this paper, I argue against the view that existing AI systems can be deemed agents comparably to human beings or other organisms. I especially focus on the criteria of *interactivity*, *autonomy*, and *adaptivity*, provided by the seminal work of Luciano Floridi and José Sanders to determine whether an artificial system can be considered an agent. I argue that the tentacles of octopuses also fit those criteria. However, I argue that octopuses’ tentacles cannot be attributed agency because their behavior can be meaningfully interpreted only in reference to the octopus’ organism as an entire system. I argue that attributing agency to AI systems faces similar difficulties, and propose an alternative characterization of these systems as responsive and contested scaffolds.

# Introduction

It is not uncommon, within Philosophy of AI, to talk of AI systems (including existing, machine-learning based AI systems) as *artificial agents*. Conceptions of agency befitting to AI systems existed since at least the early 90s (Woolridge and Jennings, 1995) and the conceptualizations of computing systems that could achieve human-like intelligence existed since their earliest development (Turing, 1950). More recently, several philosophers brought forth the idea that AI systems can possess agency in a similar way human agents can, although there are disagreements regarding whether these systems can possess some degree of agency (Coeckelbergh 2016; Johnson & Verdicchio 2019; Nyholm 2018).

I argue that existing AI systems should not be attributed agency - at least, not in the sense the term is used for human agents or even organisms. I focus on criticizing the account of artificial agency proposed by Floridi and Sanders (2004). Their account builds upon conceptions of artificial agency widespread in the AI community at the time, and proposes that AI systems, on their own, (can) possess agency, rather than as a result from their interaction with human agents. I concede that AI systems possess “quasi-agential” characteristics and propose a conception of AI (and, specifically, machine learning-based) systems that can account for the agent-like characteristics of these technologies. I introduce an analogy between AI systems and the tentacles of an octopus to lay out why AI systems might appear to be apt targets for agency attributions, but ultimately fail to achieve that goal. I instead propose that AI systems should be seen as a particular kind of *responsive* and *contested* scaffold. This characterization can maintain some of the intuitions we may have regarding AI systems without the need to wrongly attribute agency to these systems.

In **section 1**, I introduce the debate surrounding agency attribution to AI systems, with special focus on the criteria for artificial agency attribution provided by Floridi and Sanders (2004). In **section 2**, I describe the neurological structure of the octopus’ nervous system, and focus on how its tentacle can be seen as behaving independently from its central nervous system. In **section 3**, I argue that the tentacle of the octopus fits Floridi and Sanders’s criteria for determining whether an AI system counts as an (artificial) agent. I argue that we should not ascribe agency to AI systems because, in the case of the octopus’ tentacle, any kind of agency ascription must refer to the larger functional unity of the octopus as a complete system. In the same way, I argue that AI systems can be meaningfully ascribed agency only insofar as we relate the AI’s behavior to the actual agent interacting with them or the wider sociotechnical system where it behaves.

In **section 4**, I propose an alternative way of understanding the “quasi-agential” character of AI systems. I argue that AI systems should be seen as a responsive and contested kind of scaffold. They are *scaffolds*, in the sense of being a piece of the environment that an agent can interact with and modify for a given purpose. They are *responsive*, particularly sensitive to stimuli and, to some extent, capable of behavior change somewhat independently from the user. They are *contested*, in the sense that it can be quite complex to determine the ownership of the action carried out through these systems.

# 1. AI Systems as Artificial Agents: theories and critiques

There exists a longstanding debate within philosophy, robotics, and the computing sciences regarding whether AI systems could reasonably be considered agents. In the most general sense, an agent is a being with the capacity to act, and agency denotes the possession and exercise of such capacity. In philosophy, debates surrounding the nature of agency are significant within several fields. In philosophy of mind, it has been argued that agency presupposes intentionality (Davidson, 1980), free will (Ginet 1990), or consciousness (Carruthers, 2014). Alternatively, some theories of agency and agency attribution seek to establish certain requirements with regards to characteristics of their behavior (Chemero 2009).

In contrast, definitions of agency within robotics and the computing sciences proliferated in the 1990s (Woolridge and Jennings, 1995). For instance, Russel and Norvig (1995: 33) define an agent as “anything that can be viewed as perceiving its environment through sensors and acting upon that environment through effectors.” Definitions by Maes (1994), Smithers (1995), Franklin and Graesser (1996), and Christensen and Hooker (2000) share the idea that an agent can initiate and maintain interactions with its own environment, and some of them underline the importance of (self-imposed) goal-directed behavior for apt agency attribution to a system. However, definitions of (artificial) agents in this field are not set in stone, both because of the different kinds of technology that could be considered agents, and for the unclear consequences such consideration would lead to from a financial, legal and ethical viewpoint (Nwana 1996). The last decades have seen these two fields come into contact with regards to discussion of agency and agency attribution to artificial agents.

It is fascinating to note how different the kinds of requirements for agency and agency attribution are across the two fields. Partly, such differences may be derived by the different concerns that ethicists and philosophers of mind, on the one hand, and computer scientists on the other may have when it comes to the definition of what an agent is. However, since the early 2000, those two fields have gotten in increasing dialogue with one another - a dialogue clearly signaled by a particularly prevalent account for which AI systems can be legitimately considered agents.

## Floridi and Sanders’ Criteria for (Artificial) Agency Attribution

One of the most influential accounts of agency in the context of philosophy of AI and computing is the one proposed by Floridi and Sanders (2004). They provide a definition of agency applicable to AI systems without presupposing the presence of sophisticated psychological capacities such as intentionality, consciousness, or free will, in order to tackle the problematization of responsibility attribution in human-technology interaction. According to the authors, and consistently with computer and software science debates on the same issue at the time, a piece of technology can be considered an artificial agent insofar as:

1. The technology is *interactive*: it can both influence and be influenced by its environment and events that occur therein;
2. The technology is autonomous:[[1]](#footnote-0) it can operate out of its own initiative rather than exclusively in reaction to events in its environment;
3. The technology is adaptive: it can learn from previous interactions and adjust its behavioral output to optimize it.

Consider a self-driving car equipped with an AI system to determine the quickest route to a destination inputted by the driver. Let us imagine that, while navigating, the AI system would be fed information regarding an obstacle in the originally planned itinerary. The AI system then changes itinerary accordingly without informing the driver. Here, the AI system underlying the car’s change in route is *interactive*, i.e., sensitive to relevant stimuli and capable of acting upon them; *autonomous*, insofar as the route change (the behavior change) occurs internally to the AI system without input from another entity; and *adaptive*, as it changes its behavior in order to make its behavioral output match with the given task with relative independence and specifically thanks to its own capacities. According to Floridi and Sanders’ view, the AI system can, in other words, be rightfully considered agents - i.e., they can be rightfully attributed agency.[[2]](#footnote-1)

It is important, here, to clearly distinguish the sense by which Floridi and Sanders understand agency in contrast to other uses of the term applied to artifacts. Accounts of artificial agency proposed by Verbeek (2011) and Latour (1996) view it as a (however special) sort of causal efficiency. In contrast, Floridi and Sanders intend to distinguish “artificial agents” from other kinds of artifacts in virtue of some of their intrinsic characteristics. Their account of artificial agency also differs from Johnson and Verdicchio’s (2019) account of triadic agency, or Nyholm’s (2018) account of collaborative agency. While neither account views AI systems as (moral) agents, they intend to identify a phenomenon emerging from human-AI interaction irreducible to human agency alone. In contrast, Floridi and Sanders’s concept is meant to identify artificial agents, rather than a kind of agency emerging from coupled human-AI systems.

Importantly, Floridi and Sanders’ criteria to determine whether an artificial system qualifies for apt agency attribution are internalist, rather than externalist. Specifically, the authors are claiming that an artificial system that fits their criteria are not simply apt targets for agency attribution: rather, having the capacities they list *makes* these systems agents. Their qualification as agents do not rest on our own possibility to interpret them as agents (in an arguably Dennettian fashion), nor on justifying pragmatic or “folk” usages of the term “agent” towards artificial entities. The criteria they provide are meant to define what an artificial agent is based on their internal properties.

Floridi and Sanders’ account can be seen as a particularly prominent philosophical conceptualization of AI systems as a kind of agent, working as a conceptual bridge between the computing sciences and philosophy. The following discourse surrounding the idea of “AI agency” has significantly contributed to a significant amount of philosophical research surrounding the implications of viewing AI as having, at the very least, agent-like features. On the one hand, there is a significant amount of theories research surrounding the concepts of “fair AI” (Dwork et al., 2012), “trustworthy AI” (HLEG, 2019) or “responsible AI” (Dignum, 2019), which attribute to AI systems qualities that would typically be characteristic of (human) agents.[[3]](#footnote-2) On the other hand, other AI ethicists have explicitly endorsed a view like Floridi and Sanders’ to discuss the possibility of designing AI systems as “artificial moral agents” (AMAs) capable of moral behavior (or of at least avoiding immoral one) through the implementation of moral principles (Himma, 2009; Formosa and Ryan, 2021) or even virtues (Stenseke, 2023) in their design.

## 1.2 Criticisms of Agency Attribution to AI Systems

As many promoters of the idea of artificial agency acknowledge (Floridi and Sanders, 2004; Floridi, 2023; van de Poel, 2020), AI systems do not possess many of the characteristics and capacities of developed human agents – such as, intentionality, free will, or self-consciousness. As mentioned above, many philosophical accounts of agency build on these concepts in order to define the (minimal) requirements for apt agency attribution. Given the gap between traditionally conceived requirements for agency attribution and the nature of contemporary AI systems, to argue that AI systems can be attributed agency, one must view those traditional conceptions of agency as insufficient. It is exactly the perceived inadequacy of traditional criteria for agency, combined with the desire to account for the distinctive nature of AI systems, that led many philosophers to come up with criteria to capture what non-human agency would look like.

One of the more noteworthy criticisms of treating AI as a kind of agent was brought forth by Bender et al. (2021). Somewhat echoing Searle’s (1980) “Chinese Room Argument”, the authors argue that AI systems are capable of processing information, but are not actually capable of actually understanding the information they are handling. Machine learning-based AI systems are programmed through “training”, being fed an input until an output deemed apt by the programmer is produced. While the AI system can then be capable of solving given tasks much more efficiently than human beings, the operations computed by the AI system are: 1) mirroring the same result of its training period as much as possible; and 2) not understanding of the problem it learnt to solve and of whether it is repeating something incorrect, out of context, or socially inappropriate.

While AI systems may not be considered agents in the same way human beings are, there is a group of more forgiving accounts of agency that AI systems may satisfy. Non-representational accounts of agency attribution do not require the presence for complex mental states or representations in the supposed agent. For instance, Barandian et al. (2009) argue that a system may be ascribed “minimal” agency insofar as: 1) the system is distinguishable from its environment; 2) it is an active source of interactions with(in) that environment; and 3) it does so in accordance to a certain goal or norm generated from within the system itself. Such an account of agency, similarly to that of Floridi ad Sanders, do not include the presence of mental representations as a requirement.[[4]](#footnote-3)

However, these accounts are still insufficient for AI systems to be attributed agency. As Jaeger (2023) argues, the agency that characterizes biological systems (even the simplest ones, such as bacteria) not only originates within their own organization (as all non-representationalist accounts of agency agree upon in some measure). Rather, such behavior is always aimed at the carrying out of the biological processes of the organism, the self-manufacturing and self-determination[[5]](#footnote-4) that ensure its persistence. While AI systems may be capable of great computational complexity, they are not capable of self-determination: their behavior is not aimed at self-maintenance and persistence.[[6]](#footnote-5) Furthermore, the “operational domain” of AI systems (i.e., the kinds of entities that are capable of recognizing within a given environment” are typically more limited in kind compared to more basic biological agents. A generative AI may be able to interact with and produce language outputs, and a sophisticated recommendation system can read and potentially anticipate people’s preferences - however, they are typically limited to one kind of (informational) entities or stimuli in their operation, compared to living beings. In this sense, while they may be technically interactive, such interactivity is comparatively limited.

# 2. The Octopus and Its Tentacle: A De-Structured Cognitive System?

The octopus is a fascinating creature, due to its peculiar neural structure and embodied intelligence. There have been several observations of intelligent use of and interaction with environmental resources, including the use of tools (Finn et al., 2009), the possibility to solve puzzles such as opening jars (Fiorito et al., 1990) or to orient themselves within complex environments (Hvorecny et al., 2007). While octopuses are typically solitary (Mather 2019), there have also been observations of complex social behavior (Godfrey-Smith, 2016), and they appear to be capable to identify and remember other individuals, including both octopuses (Tricarico et al., 2011) and human beings (Anderson et al., 2010). Part of what makes these animals unique is the combination of their biological capacities, such as the elasticity of their tentacles combined with the presence of suckers that can support both prehensile and sensory functions, and the peculiar distribution of their nervous system. This peculiar combination of high neural concentration within the tentacles and the sensorimotor complexity of the tentacles themselves is the primary grounds for the overview of research findings and theories I provide in this section.

The tentacle of the octopus has been described as “curiously divorced” from the animal’s central nervous system (Hanlon and Messenger, 2018: 29). In a primary sense, such characterization is grounded in the specific neurological structure of the tentacle. As mentioned above, 3/5th of the octopus’ nervous system is distributed within the tentacles. This means that neural connections within the tentacles and between tentacles and central nervous system are particularly dense. Furthermore, it appears that proprioceptive information from the arms does not reach the central nervous system (Graziadei, 1971), and whenever the central brain issues motor commands to the tentacles, the communication reaches multiple tentacles, but only one carries out the relevant action (Zullo et al., 2009). Furthermore, the tentacle’s nervous structure houses ``stereotypic motor programmes”, motor patterns that store the spatial details of movements that are commonly naturally occurring – likely acquired over time through reiterated interaction with the environment (Sumbre et al., 2001). The octopus’ tentacle exhibits other significant behavioral peculiarities. It was found that the tentacle retains its sensorimotor capacities and reflexes up to 3 hours after being severed (Fraser Rowell, 1963). The tentacle itself is capable of flexible and precise movement with the function of sensorimotor and chemical inquiry of different pieces of the octopus’ environment, as well as the movement of the octopus within that environment and the handling of items in different ways.

Richter et al. (2015) studied the neurological and behavioral characteristics of octopuses fetching an object, i.e., grasping it and bringing it to the animal’s mouth. In order to grasp an object, the tentacle needs to bend at specific points in order to not under- or over-reach the object in question – which, given the lack of a bone structure, is rather computationally complex. However, the point at which the tentacle bends is not determined by the brain, but within the tentacle’s own neural structure. Because of the sensitivity of the suckers, which process chemical information about what they touch, the object’s potential role for the octopus can be identified quickly. The tentacle’s movement – which is determined internally rather than in communication with the brain – can vary depending on what the octopus wants to do with the object in question.

The neurobiological independence of the tentacle from the central nervous system has led some researchers to view the tentacle as a separate cognitive system, with even capabilities for possessing a separate consciousness from the central nervous system (Carls-Diamante, 2022). While such a claim is far from uncontested (van Woerkum, 2020), the octopus can be seen, at least, as a de-structured cognitive system. Its tentacles have a degree of neuro-behavioral flexibility and independence incomparable to the organs of other animals. Granted, the tentacle is not capable of self-sustenance (for that is a capacity of the octopus as a whole), and while it is the locus of internalization for motor patterns, tactile and chemical information about items the octopus interacts with are memorized within the animal’s central brain. However, the tentacle is adaptive insofar as sensorimotor coordination goes, thanks to its complex musculature, its sensitivity to chemical and tactile stimuli, and its neurologically enclosed structure.

# 3. From AI to Octopi and Back

I now introduce an analogy between machine-learning based AI – which qualify, according to many, artificial agents – and the tentacles of the octopus. The analogy is primarily meant to question the idea that these technologies *should* be ascribed agency in any meaningful way. However, the analogy is meant to grant that AI systems have some characteristics that differentiate them from other kinds of technologies - something that many views skeptical of “AI agency” may tend to underplay.

I will talk of the tentacle in the same way that Floridi and Sanders talk of AI systems - as if they were agents. However, I wish to pursue an opposite goal. By drawing the analogy between AI systems and tentacles, I argue that we should not ascribe agency to AI systems. I first apply the kind of language typical of agency to tentacles; then argue why we should not use that kind of language, or agency ascription, to the tentacles; and, finally, that the AI system, similarly to the tentacles, should not be ascribed agency, either.

## 3.1 The analogy between tentacles and AI systems

Let us now turn to see how the octopus’s tentacle fares with Floridi and Sanders’s criteria. Firstly, the tentacle is *interactive*: it is a system that acts upon and is acted upon by its environment. Thanks to its complex neural and sensory structure, the tentacle is very receptive to environmental stimuli and can differentiate between different kinds (e.g., food vs non-food tokens) presented to it. The tentacle can also react to those stimuli without communication with the central nervous system. The refined musculature and flexibility of the tentacle enable a great variety of potential movements – it affords several ways to interact with its environment. Its sensory apparatus enables rather sophisticated tactile and chemical information processing, enabling complex behavior without communication with the brain.

Secondly, the tentacle is *autonomous*: its internal states change independently of external input and behave independently of such input. The octopus’ tentacles can react to stimuli independently of their central nervous systems. It was claimed that motor patterns acquired from interaction with the environment are stored within the tentacle’s neural structures, rather than the brain. The tentacle can also behave without coordination with other tentacles or with other sensory apparatuses, and even when actually detached from the rest of the body. In other words: the tentacle has a surprisingly degree of autonomy, as its behavior appears to be initiated from within itself, at least when it comes to neural-based processes (perception, information processing, and movement). This does not mean that the octopus’ brain has no control over the tentacle’s behavior. Rather, when it comes to sensorimotor, tactile, and chemical information processing, the tentacle behaves with significant independence from the central nervous system.

Finally, the tentacle is *adaptive*. This is especially clear in the fetching experiments from Richter et al. (2015), where the tentacles were observed to follow specific patterns based on the overall movement function. The hypothesized storage of motor patterns learned by the octopus and stored within its tentacle rather than its brain would suggest that the memorization and enactment of such patterns is carried out by the tentacle. The motor patterns that the tentacle actually enacts are, most likely, acquired through interaction with the environment and memorized within the tentacle’s neural structure rather than in the brain. The tentacle’s purpose is primarily orientation and tactile and chemical exploration, activities presupposing the possession and enactment of motor patterns. When it comes to actions such as fetching, or the sensory exploration of items, it appears that the tentacle is itself sufficiently adaptive with regards to the aim of such activities.

This analysis leads to a clear conclusion: the octopus’s tentacle fits the criteria provided by Floridi and Sanders. The tentacle is a surprisingly complex structure, capable of identifying diverse kinds of stimuli and reacting to them in a variety of ways, in virtue of its neurobiological and muscular structure. Of course, its capacities for decoding information and reacting accordingly are limited to tactile and chemical information. However, such capacity, despite its limited scope, entails a degree of computational complexity that is utterly incomparable to that of human perception.

Machine learning-based AI systems are not particularly different from the tentacle in these regards. While their capacities are limited to processing rather specific kinds of information, their sophistication in decoding and identifying relevant information is far superior to that of human beings. Their behavior is prominently carried out without immediate oversight or input from human agents, and they adjust such behavior to solve the task at hand more efficiently. Overall, it is apparent that, in terms of interactivity, autonomy and adaptivity, AI systems and the octopus’ tentacle stand on the same grounds. It is then apparent that, to attribute agency to AI systems, in accordance with Floridi and Sanders’ account, one must be willing to attribute agency to the tentacle, too. However, there is a significant reason for not attributing agency to neither the tentacle nor to machine learning-based AI system – which is what I call the lack of *functional unity*.

Let us return to the fetching experiments carried out by Richter et al. (2015). Let us assume that the fetching movement carried out by the octopus is, as many in the literature seem to endorse, a piece of behavior carried out entirely by the tentacle, without involvement from the central nervous system. We have the opportunity to observe the behavior being carried out by the tentacle in relation to the handled item and the rest of the organism of the octopus: the tentacle is, after all, attached to the animal, regardless of its neurobiological independence. The authors denominate such behavior as fetching: the tentacle grabs the food item, and brings it *where the mouth would be*. This specification is crucial – not so much for the piece of behavior itself, but rather *for how we are able to interpret the behavior in question*. Suppose that the tentacle would be detached by the main body, and would handle the food item with an identical movement – and with identical neurological activity from within the tentacle. We would observe the tentacle grab the food item and bend itself, potentially even releasing the item. But what would enable us to identify that piece of behavior as *fetching* rather than, say, mere inspection? Any of these kinds of action can be attributed to the tentacle – or we can understand what the tentacle is “doing”, and make sense of its behavior as meaningful – only insofar as we can observe the behavior of the octopus as a whole. Otherwise, the tentacle’s behavior might simply be described as purposeless twitching. It is in this sense that the attribution of agency to the tentacle is only possible when keeping into account the behavior’s *function for the whole organism.* Just as a piece of behavior is describable as “fetching” thanks to the food being brought to the animal’s mouth, it is by observing the entirety of the animal’s behavior that we can make sense of what the tentacle is doing. The issue is not with any of the tentacle’s internal properties, but with how we are capable of attributing agency to it. The problem is with our (im)possibility to interpret the tentacle’s behavior as an action and the tentacle as an agent. Hence, interactivity, autonomy, and adaptivity are not sufficient to make the tentacle an apt target of agency attribution.

AI systems can be said to be in the same position: their behavior can be described in agential terms only insofar as we refer to the agent that is interacting with the system itself. One can rightfully claim that AI systems are interactive, autonomous, and adaptive, in accordance with Floridi and Sanders’ criteria. However, in order to understand the behavior of an AI system in terms that go beyond the language of information processing, one has to take into account the AI’s wider context and the people that are interacting with the system, and that the system is processing information for. The AI system of a self-driving car is “driving” only insofar as the car’s user decides where to go and provides the system with relevant information. Without taking into account the driver, the AI system is processing information about the optimal route, inputting directions to the car and potentially changing the given route. However, the AI is not driving the car; or rather, these kinds of “micro-decisions” may be subscribed under the activity of driving only insofar as one subsumes that activity under the wider practice of driving, which is decided upon, initiated and concluded by the driver. A recommender system only works in relation to a consumer that desires certain products or social connections, and its behavior is only understandable as a fulfillment of an action initiated by the consumer. AI systems cannot be aptly attributed agency on their own because their behavior is understandable as an action only insofar as we take into account the user of the AI system, and take the user to be initiator of the behavior in question. For the AI system to be interactive, autonomous and adaptive is not enough to take it as a fitting target of agency ascription, for its behavior can be described as a purposeful action only insofar as we take it to be part of a larger system, just as it is the case of the tentacle of the octopus.

It might be noted that, in explicit contrast to Floridi and Sanders (ibidem, 365) I am claiming that for an entity to be considered an agent, that entity requires, in some sense, intentionality. The piece of behavior in question (the tentacle twitching) may be considered as resulting from an agent’s (i.e., the octopus) behavior (e.g., moving away from a stimulus) only insofar as the entity which is considered an agent has some kind of *volitional* or *intentional* relation with the item actually producing the behavior (i.e., the tentacle). Floridi and Sanders disagree that such a specifically mental relation would necessarily be a requirement for an entity to be considered an agent; furthermore, such a requirement would automatically preclude artificial systems from being agents. In this sense, I would like to be more precise: functional unity, more so than intentionality or volition, entails the self-determination of goal-directed behavior. As Floridi and Sanders’ point out, artificial systems can be designed to be goal directed; however, and similarly to the tentacle, they are not capable of generating their own goals and, at the same time, pursue those goals. For this reason, neither the octopus’s tentacle (which can behave in a goal-directed way), nor the octopus’s brain (which can arguably generate their own goals) can be considered valid targets of agency attribution - and neither possess what I call functional unity. Theoretically, functional unity so understood does not require volition or intentionality.

## 3.2 Objections and counterarguments

There are several potential objections to my argument that I want to address here. Firstly, it might be objected that the characterization of AI systems as agents precisely makes sense within a wider sociotechnical system. After all, Floridi introduces the idea of artificial agents precisely within his theory of “infraethics”, according to which morally irrelevant actions of an agent or an artifact can acquire moral significance at a systemic level. In this view a technology’s (piece of) behavior is morally relevant only when considering its role from a wider, zoomed-out perspective. A self-driving car may “decide” for a route change that may later lead to an accident. Hence, a piece of morally irrelevant behavior may acquire moral significance when taking the perspective of the entire sequence of events and people involved therein.

The problem with this objection is simple: the polyvalent moral (in)significance of a piece of behavior (be it from a technology, a human being, or what have you) does not depend on whether the behaving entity ought to be considered an agent or not. What agency attribution to AI systems provides, according to Floridi and Sanders, is the idea that these systems can be targets of accountability attributions. However, there are ways to understand and address the moral significance of these (and other) technologies' behavior, even taking an “infraethics” perspective, without viewing these technologies to be apt targets of agency (or accountability) attribution.[[7]](#footnote-6)

The second objection is already mentioned in Floridi and Sanders’s argument. AI systems are clearly not agents in the same way that human agents are. Hence, the criteria they provide are not meant to be requirements for agents, but specifically for *artificial* agents – whereby, while still capable of being attributed some forms of responsibility attribution, are a different kind than natural agents.

To this objection, I feel compelled to ask how far one is willing to use a concept, before that concept loses its original meaning. AI systems, just as the tentacle of the octopus, fail to meet any “traditional” philosophical requirement for agency ascription, including both representational and non-representational ones. If they fail to meet *any* existing criteria for apt agency ascription, then attributing a “special” kind of agency to AI systems can only be possible when adding a set of asterisks to whatever “special” understanding of agency that ascription is supposed to entail. There has to be legitimate motivations to make such a theoretical move – in this case, arguably, the specific nature of artificial agents as possessing a degree of initiative and independence. Even if using the term more loosely would be fine in folk use, where a lack of conceptual exactness may be more frequent, such confusion should be avoided when the term (or the assumption of the aptness of the term, as in the case of “trustworthy” or “fair” AI guidelines) is used in expert settings. However, as I have argued, agency ascription requires some functional unity (which is missing in the case of the tentacle, and in the case of the AI system). Hence, in my view, to attribute a special kind of agency to artificial systems in order to account for certain intuitions we might have about them would ultimately alter the concept of agency to the point of describing something radically different to what it is typically used for. What I am criticizing is the (mis)use of an existing category to subscribe to a new kind of entity, at the cost of making such an exception to make room for our intuition that the original category loses its significance. If AI systems are just like agents, except that they are not really like agents but something different, then it would be much more apt to come up with a distinctive category for these entities.

My objection could be countered by underlining a difference between traditional accounts of (natural) agency and accounts of artificial agency. Traditional accounts inquire the intrinsic conditions of what defines agency as such. To claim that agency occurs thanks to mental causation, intentionality, or similar mental processes requires defining an internal, non-visible[[8]](#footnote-7) condition of the agent that makes their behavior describable as that of an agent. It is a top-down approach: to identify what kind of thing qualifies for agency ascription, one must define the agent’s intrinsic characteristics. In contrast, Floridi and Sanders’ approach is bottom-up: the qualification for apt agency ascription depends on the manifest behavior of the agent, rather than any internal qualification. While agency ascriptions of a natural (i.e., human) agent involve identifying conditions that are internal to the agent – be it intentionality, mental causation, or even self-organization – that does not apply to Floridi and Sanders’ criteria. It might be argued that by avoiding identifying internal criteria for agents, their account would actually be open for the eventuality that an artificial system might actually develop the characteristics of natural agents, rather than possessing them from the start.

While there are legitimate merits to this non-internalist approach to agency ascription, the objection falls short of its goal for two reasons. Firstly, the merit of these approaches to agency can make room for the *eventual* development of artificial systems to the point that they can be aptly ascribed such quality. However, they were also meant to apply to artificial systems that exist right now - and the specific criteria they propose do not suffice for apt agency attribution, as I have argued above. Artificial systems are presently being talked of as agents or possessing agent-like qualities. In other words, Floridi and Sanders’ account, while having a theoretical advantage over internalist views of agency, is still not adequate to describe *actual* supposed artificial agents (which is, admittedly, what the authors intend it to do) without denaturalizing what the concept of agency is supposed to define.

A final objection may involve the idea of functional unity that grounds my argument against agency attribution to the tentacle. It might be argued that, when people partake into a social practice that requires them to act in specific ways, their behavior is fully understandable only in reference to the wider practice they partake in. For instance, on a ship, different sailors may have different tasks, but while their behavior can be singled out, one can fully make sense of each of the sailors’ behavior only when taking into account the entire social practice of navigation, which is brought about collectively. However, compare this case of navigation with one of “orientation”, where a person utilizes a compass and their behavior can only be understood when understanding how the compass works. The behavior we observe in the compass’s user can be described in agential terms despite the fact we do not possess a complete overview over its meaning. The same can be said for a more complex practice such as navigation on a ship. While the steering, movement, calculations, order-giving and decision-making of each of the sailors can be fully made sense of, only insofar as the whole social practice can be taken into account, each of these actions can be described in intentional terms even when looking at them in isolation. The same cannot be said for the compass’s behavior, nor for that of AI systems, whose activity can be identified as purposeful or goal-directed only insofar as we keep their users in mind.[[9]](#footnote-8) My account strongly differs from accounts of “collaborative agency” to explain responsibility attribution in the context of human-AI interaction (Nyholm, 2018; Johnson and Verdicchio, 2019). According to this view, AI systems, while not responsible agents, are still capable of some kind of basic agency entailing the pursuit of goals in a way that is sensitive to the representation of its environment. However, such goals are not generated from the AI system itself but by its programmer and/or its users, in the same way that the tentacle does not have any goals of its own except for those of the octopus.

Overall, I have defended my argument against several potential objections, and further elaborated upon why AI systems should not be ascribed agency. One should not attribute agency to AI systems because 1) they do not meet any criteria for other kinds that would normally be categorized as agents; 2) existing criteria for determining whether AI systems count as agents also apply to entities that are not agents; and 3) attributing agency or characteristics typical of agents to AI systems leads to several problems surrounding the proper attribution of responsibility, fairness, and trustworthiness. I now turn to an alternative characterization of AI systems that can grant them the seemingly agential characteristics that Floridi and Sanders try to ascribe them without mistakenly characterizing them as agents.

# 4. An alternative to artificial agency: AI systems as responsive and contested scaffolds

Accounts proposing criteria for agency attribution to AI systems, such as Floridi and Sanders’s are grounded on the need to characterize them as a specific kind of technology, radically different than tools like hammers, rotary phones, or cars are. I sympathize with such a motivation. AI-based technologies are capable of changing their behavior in context-responsive ways with a potentially relevant degree of independence from the user, and in potentially unforeseen ways. I now propose an alternative characterization for these technologies, meant to bypass their alternative characterization as either genuine agents or stochastic parrots, by defining AI systems as a distinct kind of *scaffold*.

The term “scaffolding” has become significant in philosophy of mind, and specifically for theories of embodied, embedded, enactive and extended cognition. These theories share the common assumption that cognitive processes are enabled or potentially co-constituted by the body and (parts of) the social and material environment of the agent - with intradisciplinary disagreements regarding the manner and extent of such expansion (Newen et al., 2018). Within this framework, Sterelny (2010; see also Varga, 2018) introduced the theory of scaffolded cognition. Framing his argument in the idea of niche-construction, the environmental modifications brought about by organisms that can alter their evolutionary trajectory in the long run (Odling Smee et al., 2003; Matthews et al., 2014), Sterelny argues that human beings are capable of scaffolding their cognitive processes through and within parts of their environment. Such scaffolding has the potential of not just facilitating cognitive processes, but also to potentially enable new processes and practices that would have been unachievable without those environmental modifications. The term scaffolding, here, entails the construction and interaction with a (somewhat temporary) structure in order to facilitate or enable cognition to be carried out. Other authors claim that affective processes, which include emotions, feelings, and moods, can also be scaffolded (Slaby, 2016; Coninx and Stephan, 2021). The construction and use of these structures, which can be both material or cultural (e.g, social practices) is characteristic of (albeit not exclusive to) our species, and it has been argued that it significantly contributed to the development of cultural inheritance within the *Homo Sapiens* genus (Sterelny, 2012). The term scaffolding indicates the process of manipulation and appropriation of pieces of the environment to carry out a given practice or process; a scaffold is such a piece of the environment, which can be both an object, a person, or a practice.

I propose that AI systems should be considered a particular kind of scaffold, a piece of social and material infrastructure human beings interact with to offload and enable certain cognitive, affective and decision-making practices. Many kinds of artifacts have been analyzed through a scaffolded mind approach, including AI-based ones (Hernandez-Orallo and Vold, 2019, e.g.). I believe it is important to look at AI systems specifically and to provide them with two characteristics that make them distinctive in contrast to other kinds of artifacts. I specifically argue that AI systems are a kind of *responsive* and *contested* scaffold, with some agent-like characteristics but not apt entities for agency attribution

With the term *responsive*, I mean to capture some of the characteristics of AI systems as interactive and adaptive. AI systems are particularly receptive to the environment and the relevant information therein, so much so that they can change their behavior depending on it. A self-driving car is receptive (thanks to sensors and cameras) to specific events in its environment and to general information about traffic it may be able to acquire through, e.g., Internet connection. Not only does it possess an apparatus for gathering the relevant information for the task(s) it is meant to fulfill. By processing this information, the AI system is capable of behavior change without additional input from the agent. One may say that the self-driving car’s AI system is processing information and adapting its behavior thanks to their responsiveness – something a normal car is not capable of. However, the driving is carried out by the human sitting in the car; the initial identification of what kind of information counts as relevant for the car, including the recognition of pedestrians, is also originally inputted by the car’s designers – although the development of computations for these tasks is developed by the AI system itself. While one may not attribute agency to the AI system (or the car), it is undeniable that the system is responsive, both receptive to the relevant information it is intended to process and capable of changing behavior in response to the information thereby processed.

Secondly, and potentially more importantly, I define AI systems as a kind of *contested* scaffold. The contested nature of AI systems entails two different kinds of potential issues from the perspective of people interacting with them. AI systems behave with a relative degree of independence from the human agents interacting with them. The self-driving car can make a change in the route to optimize the travel time, without input from the driver with regards to that behavior change. There is a clear sense by which AI systems are a part of technologies that are used by human agents: people use self-driving cars, doctors use medical equipment with machine learning technology, social media users create profiles and consume content that is provided by recommender systems. However, there exist multiple senses in which the ownership of these technologies is contested. This is partly due to the responsiveness of AI systems: because of their changes in behavior for the purposes of task completion, people are not always in complete control over the AI system’s behavior, at least when it comes to “microdecisions”. Furthermore, AI systems are intrinsically opaque technologies. The high degree of complexity of their structure and computations makes the AI systems opaque to the human actors interacting with them (including their designers) (Castelvecchi, 2016; von Eschenbach, 2021). While a person using an artifact ought to be considered the subject of the actions they carry out with it, in the case of AI systems this picture may get quite more muddled, because the control and understanding people have of AI systems is intrinsically limited. In a first sense of the term, AI systems are contested scaffolds because the extent by which people can relevantly[[10]](#footnote-9) interact with, control, or understand them is questionable.

There is a secondary sense in which AI systems are contested scaffolds, i.e., with regards to the involvement of multiple human agents with this kind of technology. Take a social media site’s recommender system. The recommender system has two purposes depending on the relevant user perspective one takes. From the perspective of a social media site user, the purpose of the recommender system is to provide them with content they want to engage in – be it posts from their social ties or other online actors they follow, or recommendations for further social ties. From the perspective of the service provider, however, the recommender system fulfills a different function: by becoming more and more apt in its recommendation to social media users, it provides the platform’s service providers with information regarding its users – information they may then use and sell to third parties. Here, what is contested is not simply the origin of the behavior (whether, and in what regards, the relevant behavior originates from the AI system or a human agent). Rather, it is not clear who is the relevant human agent the AI system is working *for* – i.e., who is the actual owner of the processes scaffolded by the AI system. This unclarity in ownership may appear in many different implementations of AI systems. This is especially so because, given the relative opacity of these technologies in their adaptivity and behavior, it is essential for its designers to gather further data to, at the very least, maintain an overview over the AI’s behavior, if not adjusting it over time for the purposes of optimization. Contested ownership is, hence, especially clear when the given AI system may be considered owned both by the user it is designed for and the company that produced it. This dynamic, which is already discussed much in ethics of the scaffolded mind and its related theories (Rowlands, 2009; De Presteer, 2011; Gallagher, 2013) has led some authors to argue that we should regard the relevant technology as being co-owned by both relevant actors (Vold and Liao, forthcoming). However, the extent and significance of the different kinds of claims over AI systems can be complex to assess, despite its importance in determining the claims and rights of different stakeholders over the technology and its functioning. In this sense, then, the contested nature of AI systems as scaffolds also entails a problematization of whose activity is the technology supposed to enhance or enable.

Of course, our construction of and interaction with scaffolds in general should not generally be seen as perfectly “clean”, without uncertainties and risks; in fact, many scaffolds can become (or even be designed to be) maladaptive (Timms & Spurrett, 2023). There are three things to note here. Firstly, it is noteworthy that the origination of the action itself is contested in the case of AI systems, rather than a variety of other factors that may characterize our relationship with a scaffold (e.g. its fragility or persistence, the way our affectivity may change towards it over time, how we may become dependent on it…). Secondly, their contestedness does not simply entail a lack of control from a user of the scaffold, but specifically a tension between more than one agent with regards to who (or what) originated the action in question. Finally (although this may not, in principle, be exclusive to AI systems) such contestedness appears to be an intrinsic and unavoidable feature of AI systems.

# 5. Concluding Remarks

In this chapter, I have argued that AI systems should be attributed agency to the same extent that the tentacles of octopuses should: not at all. After explicating the reasoning for this comparison, I have argued that some of the most prominent criteria for attributing agency to AI systems, provided by the seminal work of Floridi and Sanders (2004) can also be aptly applied to the tentacles of octopuses. However, in the same ways the tentacles can only be attributed some degree of agent-like characteristics, even when severed, insofar as their behavior is explainable in reference to the entire organism, so lack AI systems the kind of functional unity that is typical of entities we typically deem as agents. I then proposed an alternative characterization for AI systems, as responsive and contested scaffolds, so as to capture the reasons why we may take these technologies to be agents without misattributing them agency.

There is one particular issue that I wish to address, here: would it be possible, in theory, for an AI system to acquire functional unity and be, according to my account, an apt target for agency attribution?[[11]](#footnote-10) In theory, I do not see why that would be an impossibility. I do not believe that existing AI systems, even those as complex as generative AIs, would qualify as agents according to the argument I have proposed In the eventuality of the emergence of a so called “General” AI, it would be possible for such an AI to behave with the kind of functional unity that, I argue, is necessary for necessary agency attribution, although it is not entirely clear when such an AI would be successfully develop.. What might be more closely achievable, given the current state of development of AI, might be the construction of several AI systems to coordinate in such a way that, *when considered together*, would be able to behave as a functional unity. Similarly to the relative independence of the octopus’ tentacles from its central nervous system, a set of AI systems that would technically be semi-independent from one another could be programmed to work together as a functional unit. However, such a conglomeration of AI systems would be the apt target of agency attribution, rather than the individual AI systems. And in this sense, my argument with regards to individual AI systems would stand.

While my characterization of AI systems as responsive and contested scaffolds is novel, the manner it would frame out philosophical and ethical understanding of AI systems fits with many existing philosophical perspectives in the field. For instance, many of the issues regarding the contested ownership of AI systems are dealt with in theories of meaningful human control (Santoni de Sio et al., 2023). My aim was to single out a particularly reoccurring aspect of the academic and political discourse surrounding AI systems, that is their (more or less implicit) characterization of agents, and to argue that such attribution is unjustified. However, I also argued that these technologies are particularly specific compared to others: hence, while I argued that they should be characterized as scaffolds rather than agents, in my view, they should be characterized as a kind of responsive and contested kind of scaffold. This characterization can encapsulate the agent-like characteristics that many recognize, without making the category mistake of attributing them agency.

The characterization of AI systems as scaffolds, rather than agents, would not just be a solid alternative to agency ascription to AI systems, in spite (and partly accounting for) some of their characteristics. A significant consequence of such conceptualization would be, in line with Ryan (2020) and Conradie and Nagel (Forthcoming), that we should drop the ascription of characteristics such as trustworthiness, fairness or accountability to AI systems. These terms are pretty widespread within AI ethics even without explicitly drawing from their characterization as agents – even though these are very clearly characteristics typical of agents. While in the case of trust and trustworthiness there may be a way to relevantly distinguish whether it regards agents or artifacts (e.g., Nguyen, 2022), that may not be the case with fairness, responsibility or accountability. While the meaning of these terms may be intended differently in AI ethics, depending on whether they refer to human agents or to AI systems (as it is the case with fairness: see e.g., Mulligan et al., 2019). To be fair, the use of the concept of fairness may be so entrenched not just in AI ethics, but in AI and the computing sciences as a whole, that dropping it entirely may be practically impossible. However, and especially because the meaning of the concept within those fields tends to have a quite different meaning than when the concept is used in philosophy, researchers should have the responsibility of making such difference as clear as possible. As pointed out by Kulis (this volume), the non-exact application of certain concepts to given entities, while sometimes unavoidable (especially in everyday language) not only brings to light our assumptions and expectations towards that entity. In his view, attempting to translate the characteristics and capabilities of AI through attributes typical of human beings (such as *intelligence*) carries severe risks in terms of misunderstanding those entities and miscalculating our expectations out of what they can do. I very much share his concern, and view my efforts in this chapter as an expression of such concern.

# Acknowledgments

I would like to thank my colleagues at the RWTH Aachen Applied Ethics Team (especially Prof. Saskia K. Nagel, Dr. W. Jared Parmer, and Dr. Hendrik Kempt) for providing feedback for earlier version of this chapter. I also want to thank the organizers of the PT-AI 2023 Conference for providing me with the opportunity to present and share this chapter within a thriving academic community. I want to finally thank Dr. [Māris Kūlis](http://www.mariskulis.com/en/homepage/) and Dr. Nicolas Kuske for their thorough and insightful feedback, which was fundamental to extensively improve on the chapter.

**References**

Anderson, R. C., Mather, J. A., Monette, M. Q. & Zimsen, S. R. M. (2010). Octopuses (*Enteroctopus dofleini*) recognize individual humans. *Journal of Applied Animal Welfare Science, 13,* 261-272.

Barandiaran, X. E., Di Paolo, E., & Rohde, M. (2009). Defining agency: Individuality, normativity, asymmetry, and spatio-temporality in action. *Adaptive behavior*, *17*(5), 367-386.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big?🦜. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*(pp. 610-623).

Carls-Diamante, S. (2022). Where is it like to be an octopus?. *Frontiers in systems neuroscience*, *16*, 840022.

Carruthers, G. (2014). What makes us conscious of our own agency? And why the conscious versus unconscious representation distinction matters. *Frontiers in Human Neuroscience*, *8*, 434.

Castelvecchi, D. (2016). Can we open the black box of AI?. *Nature News*, *538*(7623), 20.

Christensen, W. D., & Hooker, C. A. (2000). An interactivist-constructivist approach to intelligence: self-directed anticipative learning. *Philosophical Psychology*, *13*(1), 5-45.

Coeckelbergh, M. (2016). Responsibility and the moral phenomenology of using self-driving cars. *Applied Artificial Intelligence*, *30*(8), 748-757.

Coninx, S., & Stephan, A. (2021). A taxonomy of environmentally scaffolded affectivity. *Danish Yearbook of Philosophy*, *54*(1), 38-64.

Conradie, N. H., & Nagel, S. K. (2024). No Agent in the Machine: Being Trustworthy and Responsible about AI. *Philosophy & Technology*, *37*(2), 72.

Davidson, D. (1980). *Essays on Actions and Events*, Oxford: Clarendon Press.

Dignum, V. (2019). *Responsible artificial intelligence: how to develop and use AI in a responsible way* (Vol. 2156). Cham: Springer.

De Preester, H. (2011). Technology and the body: the (im) possibilities of re-embodiment. *Foundations of science*, *16*, 119-137.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference* (pp. 214-226). https://doi.org/10.1145/2090236.2090255

de Sio, F. S., Mecacci, G., Calvert, S., Heikoop, D., Hagenzieker, M., & van Arem, B. (2023). Realising meaningful human control over automated driving systems: a multidisciplinary approach. *Minds and machines*, *33*(4), 587-611.

Finn, J. K., Tregenza, T. & Norman, M. D. (2009). Defensive tool use in a coconut-carrying octopus. *Current Biology, 19,* 1029-1030.

Fiorito, G., von Planta, C. & Scotto, P. (1990). Problem solving ability of *Octopus vulgaris* Lamarck (Mollusca, Cephalopoda). *Behavioral and Neural Biology, 53*, 217-230.

Floridi, L. (2023). AI as agency without intelligence: On ChatGPT, large language models, and other generative models. *Philosophy & Technology*, *36*(1), 15.

Floridi, L., & Sanders, J. W. (2004). On the morality of artificial agents. *Minds and machines*, *14*, 349-379.

Formosa, P., & Ryan, M. (2021). Making moral machines: why we need artificial moral agents. *AI & society*, *36*(3), 839-851.

Franklin, S., & Graesser, A. (1996). Is it an Agent, or just a Program?: A Taxonomy for Autonomous Agents. In *International workshop on agent theories, architectures, and languages* (pp. 21-35). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fraser Rowell, C. H. (1963). Excitatory and inhibitory pathways in the arm of Octopus. *Journal of Experimental Biology*, *40*(2), 257-270.

Gallagher, S. (2013). The socially extended mind. *Cognitive systems research*, *25*, 4-12.

* Ginet, C. (1990) *On Action*, Cambridge: Cambridge University Press.

Godfrey-Smith, P. (2016). *Other minds: The octopus and the evolution of intelligent life* (Vol. 325). London: William Collins.

Graziadei P (1971) The nervous system of the arms. In: Young, J.Z. (ed.), *The anatomy of the nervous system of octopus vulgaris*. Clarendon Press, Oxford, pp 45–61

Hanlon, R. T. & Messenger, J. B. (2018). *Cephalopod behavior,* 2nd ed*.* Cambridge, UK: CambridgeUniversity Press.

Hernández-Orallo, J., & Vold, K. (2019). AI extenders: the ethical and societal implications of humans cognitively extended by AI. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 507-513).

Himma, K.E. (2009) Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent?. *Ethics of Information Technology,* *11*, 19–29. <https://doi.org/10.1007/s10676-008-9167-5>

HLEG (High-Level Expert Group on Artificial Intelligence) (2019). Ethics guidelines for trustworthy AI. Retrieved from: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trust worthy-ai

Hvorecny, L. M., Grudowski, J. L., Blakeslee, C. J., Simmons, T. L., Roy, P. R., Brooks, J. A., ... & Boal, J. G. (2007). Octopuses (Octopus bimaculoides) and cuttlefishes (Sepia pharaonis, S. officinalis) can conditionally discriminate. *Animal cognition*, *10*, 449-459.

Jaeger, J. (2023). Artificial intelligence is algorithmic mimicry: why artificial" agents" are not (and won't be) proper agents. *arXiv preprint arXiv:2307.07515*.

Johnson, D.G., & Verdicchio, M. (2018). AI, agency and responsibility: the VW fraud case and beyond. *AI & SOCIETY, 34*, 639-647.

Latour, B. (1996). On actor-network theory: A few clarifications. *Soziale welt*, 369-381.

Libet, B. (1999). Do we have free will?. *Journal of consciousness studies*, *6*(8-9), 47-57.

Maes, P. (1994. Social interface agents: Acquiring competence by learning from users and other agents. In *Software Agents—Papers from the 1994 Spring Symposium (Technical Report SS–94–03)* (pp. 71-78).

Mather, J. (2019). What is in an octopus's mind?. *Animal Sentience*, *4*(26), 1.

Matthews, B., De Meester, L., Jones, C. G., Ibelings, B. W., Bouma, T. J., Nuutinen, V., ... & Odling-Smee, J. (2014). Under niche construction: an operational bridge between ecology, evolution, and ecosystem science. *Ecological Monographs*, *84*(2), 245-263.

Mossio, M., & Bich, L. (2017). What makes biological organisation teleological?. *Synthese*, *194*(4), 1089-1114.

Mulligan, D. K., Kroll, J. A., Kohli, N., & Wong, R. Y. (2019). This thing called fairness: Disciplinary confusion realizing a value in technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(CSCW), 1-36.

Newen, A., Gallagher, S., and De Bruin, L. (2018). 4E Cognition: Historical Roots, Key Concepts, and Central Issues, in A. Newen, L. De Bruin, and S. Gallagher (eds), *The Oxford Handbook of 4E Cognition*, Oxford Library of Psychology.

Nyholm, S. (2018). Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci. *Science and engineering ethics*, *24*(4), 1201-1219.

Odling-Smee, F. J., Laland, K. N., & Feldman, M. W. (2003). *Niche Construction: The Neglected Process in Evolution (MPB-37)*. Princeton University Press.

Popa, E. (2021). Human goals are constitutive of agency in artificial intelligence (AI). *Philosophy & Technology*, *34*(4), 1731-1750.

Richter, J. N., Hochner, B., & Kuba, M. J. (2015). Octopus arm movements under constrained conditions: adaptation, modification and plasticity of motor primitives. *The Journal of Experimental Biology*, *218*(7), 1069-1076.

Rowlands, M. (2009). Extended cognition and the mark of the cognitive. *Philosophical Psychology*, *22*(1), 1-19.

Russell, S., & Norvig, P. (1995). A modern, agent-oriented approach to introductory artificial intelligence. *Acm Sigart Bulletin*, *6*(2), 24-26.

Ryan, M. (2020). In AI we trust: ethics, artificial intelligence, and reliability. *Science and Engineering Ethics*, *26*(5), 2749-2767.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417-424.

Slaby, J. (2016). Mind invasion: Situated affectivity and the corporate life hack. *Frontiers in Psychology*, *7*, 174220.

Smithers, T. (1995). Are autonomous agents information processing systems?. In *The artificial life route to artificial intelligence* (pp. 123-162). Routledge.

Stenseke, J. (2023). Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*, *38*(4), 1301-1320.

Sterelny, K. (2010). Minds: extended or scaffolded?. *Phenomenology and the Cognitive Sciences*, *9*(4), 465-481.

Sterelny, K. (2012). *The evolved apprentice*. MIT press.

Sumbre, G., Gutfreund, Y., Fiorito, G., Flash, T., & Hochner, B. (2001). Control of octopus arm extension by a peripheral motor program. *Science*, *293*(5536), 1845-1848.

​​Timms, R., & Spurrett, D. (2023). Hostile Scaffolding. *Philosophical Papers*, *52*(1), 53–82. https://doi.org/10.1080/05568641.2023.2231652

Tricarico, E., Borelli, L., Gherardi, F. & Fiorito, G. (2011). I know my neighbour: Individual recognition in *Octopus vulgaris. PLoS ONE, 6,* e18710.

Turing, A.M. (1950). Computing Machinery and Intelligence. *Mind*, LIX (236), 433–460.

Van de Poel, I. (2020). Embedding values in artificial intelligence (AI) systems. *Minds and Machines*, *30*(3), 385-409.

van Woerkum, B. (2020). Distributed nervous system, disunified consciousness?: A sensorimotor integrationist account of octopus consciousness. *Journal of Consciousness Studies*, *27*(1-2), 149-172.

Varga, S. (2019). *Scaffolded minds: Integration and disintegration*. MIT Press.

Verbeek, P. P. (2011). *Moralizing technology: Understanding and designing the morality of things*. University of Chicago press.

Vold, K., & Liao, X. (2024). Neuroprosthetics, Extended Cognition, and the Problem of Ownership. In *Neuro-ProsthEthics: Ethical Implications of Applied Situated Cognition* (pp. 37-55). Berlin, Heidelberg: Springer Berlin Heidelberg.

Von Eschenbach, W. J. (2021). Transparency and the black box problem: Why we do not trust AI. *Philosophy & Technology*, *34*(4), 1607-1622.

Wooldridge, M., & Jennings, N. R. (1995). Intelligent agents: Theory and practice. *The knowledge engineering review*, *10*(2), 115-152.

Zullo, L., Sumbre, G., Agnisola, C., Flash, T., & Hochner, B. (2009). Nonsomatotopic organization of the higher motor centers in octopus. *Current Biology*, *19*(19), 1632-1636.

1. Autonomy is here understood as independence from direct control from human agents (Adams, 2001; Behdadi and Munthe, 2020). [↑](#footnote-ref-0)
2. Floridi and Sanders talk of (artificial) *agents,* rather than of *agency*. To be precise, I am interpreting their claim as follows: if a system possesses the criteria of interactivity, autonomy, and adaptivity, then it can be considered an agent *and* it can be attributed a kind of agency comparable, in some significant regard, to that of human beings or other organisms. [↑](#footnote-ref-1)
3. For some of these concepts, and most prominently that of “responsible” AI, the qualifier does not *technically* refer to a given AI system as a technology, but to the sociotechnical system and practices that revolve around its development and deployment. The reader might be asking why would then one need to talk of responsible “AI” rather than, e.g., responsible innovation. As it will become evident throughout this chapter, I share such perplexity, which is a strong motivator for my overall argument. [↑](#footnote-ref-2)
4. Arguably, existing AI systems still do not fit these criteria either, as the goals according to which they act are generated by their programmers via design, rather than from within the system itself (Popa, 2021). [↑](#footnote-ref-3)
5. A system is self-determining insofar as its behavior contributes to the establishment and maintenance of its own conditions of existence (Mossio and Bich, 2017). [↑](#footnote-ref-4)
6. It might be argued, here, that a self-driving car could be programmed with an AI system that can reprogram its route towards a gas (or electricity) station if the car is running low on fuel: in such a case, the AI system would appear capable of self-preservation. That is, however, not the case: such an AI system, being programmed to ensure that the car does not run out of fuel, is capable of preserving the car, not itself. It can preserve its functioning only insofar as it can preserve the functioning of the car. I want to thank one of my reviewers for pointing out such a potential counter-example to my point. [↑](#footnote-ref-5)
7. Even if, as pointed out by one reviewer, Floridi and Sanders’ goal with their conception of artificial agents would “simply” be to justify their claim that these systems can be held accountable, such a claim would need to rest on adequate conceptual grounds - which, in my view, are missing. [↑](#footnote-ref-6)
8. To be precise: we may recognize behavior as intentional, but we do not have the means of seeing the actual mental processes as such, at least not in everyday interaction – which is what concerns this objection and my counterargument to it. There may be an argument to be made that brain scanning tools may individuate mental processes in specific brain regions, such as the famous experiments on readiness potential (Libet 1999), but, even then, it is unclear whether one can unequivocally identify mental and neural processes (such as, e.g., readiness potential and free will). That is an issue that goes well beyond the scope of this chapter. [↑](#footnote-ref-7)
9. There may be exceptional cases where the behavior of individual human agents may not be understandable when separated from their collaborative practice or the wider social context. For instance, the art piece by Tino Sehgal presented in the 2013 Biennale di Venezia, entitled “Untitled” (2013), consisted in people looking at the art pieces in the main hall who would then flail on the ground and enunciate the name of the art piece. Here, each individual’s behavior may be non-describable in terms of agency, and it may be possible to make sense of their behavior only in reference to the art piece – i.e., the practice. In that case, importantly, part of the intention from the artist (the designer of the practice, if you will) was (likely) to puzzle the audience with the behavior of each participant of the practice. [↑](#footnote-ref-8)
10. This specification is important when comparing AI systems to other kinds of scaffolds. I may use pen and paper to perform calculations without understanding how exactly the ink sticks to the page, or the atomic processes internal to the pen. However, the extent of my understanding of the pen is sufficient for the task I am utilizing it for, and I know what to do in case it is not working as it supposedly should. That is not the case for AI systems. [↑](#footnote-ref-9)
11. I thank one of my reviewers for bringing this question forth. [↑](#footnote-ref-10)