Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

Understanding Artificial Agency

By Leonard Dung

Abstract

Which AI systems are agents? To answer this question, I propose a multidimensional account of agency. According to this account, a system's agency profile is jointly determined by its level of goal-directedness and autonomy as well as is abilities for directly impacting the surrounding world, long-term planning and acting for reasons. Rooted in extant theories of agency, this account enables fine-grained, nuanced comparative characterizations of artificial agency. I show that this account has multiple important virtues and is more informative than alternatives. More speculatively, it may help to illuminate two important emerging questions in AI ethics: 1. Can agency contribute to the moral status of non-human beings, and how? 2. When and why might AI systems exhibit power-seeking behavior and does this pose an existential risk to humanity?

Keywords

Robot rights; moral status; existential risk; reasons; language models; goals

1. Introduction

Recent progress in artificial intelligence (AI) research, especially machine learning (ML), has been staggering. Particularly large language models (LLMs) have gained public attention and captured the imagination of researchers and investors. This development creates an urgent need to reflect on these systems from both a scientific as well as an ethical perspective. Scientifically, we can ask what the capabilities, and the limits, of current state-of-the-art models are. This is also important to inform predictions about their future potential. Ethically, we can ask what

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

promises, but also what risks, these developments harbor for improving society and supporting human and animal (Hagendorff et al. 2022) flourishing.

To bridge these concerns, I propose a conceptual framework for understanding *agency* in artificial systems. As a scientifically relevant but also ethically significant property, thinking about agency can help illuminate scientifically and societally important issues regarding AI. I will propose a multi-dimensional account of agency which enables fine-grained, nuanced comparative characterizations of artificial agency. Thus, the account can serve as the basis of comparative cognition research on agency. Moreover, the framework developed here helps to clarify and make more precise the notion of agency which is central to many philosophical debates (Ferrero 2022; Schlosser 2019).

In the next section, I will propose my theory of agency. According to this dimensions account, agency can be understood along several, partially related, dimensions. In section 3, I will present six virtues of this account. To further support this account, section 4 sketches its potential for illuminating ethical questions relevant to AI agency. Section 5 concludes.

2. Dimensions of artificial agency

2.1 A multidimensional approach

My core claim is that variations in the degree of agency of different beings, specifically AI systems, can be captured along five distinct dimensions. Thus, a system's *agency profile* can be depicted as a point in a multidimensional agency space where the system's location in the space is determined by its value on each of the five dimensions. This enables fine-grained, comparative characterizations of the degree and nature of agency of different systems. This dimensional account does justice to the varied, complex nature of agency.[1] For it does not

[1] The dimensions account of artificial agency has a similar rationale to multidimensional accounts of consciousness (Birch, Schnell, and Clayton 2020; Dung and Newen 2023), causal cognition (Starzak and Gray 2021) and empathy (Newen, Griem, and Pika 2022) in animals.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

presuppose that it is always a meaningful question to ask whether system A has 'more'agency than system B. Instead, A might be more agentic in some respects and B in others. At the same time, the number of dimensions is limited to five and each dimension is assumed to be unified. Thus, the dimensional account does not depict agency as limitlessly intricate: it captures the homogeneity of agency.

Why should one not just subscribe to an account which characterizes agency as a single, unitary property, rather than through a multidimensional space? As we will see, different dimensions correspond, to some extent, to different accounts of agency from the literature. However, to realize the explanatory potential of a theory of agency, no single account – and no single dimension – is satisfactory. Due to the diverse roles the notion of agency plays, a multidimensional account is necessary to respect its explanatory promise.

2.2 The five dimensions

2.2.1 Goal-directedness

In this sub-section, I will describe the five dimensions which – according to my dimensions account – jointly fully characterize agency. The first dimension is *goal-directedness*. The literature in different disciplines – philosophy, psychology and AI – agrees that agency has something to do with the capacity to pursue goals (e.g. Glock 2019; Kenton et al. 2022; Liljeholm 2021; Orseau, McGill, and Legg 2018). Interchangeably, one can say that agents have preferences or wants and sometimes behave based on them. That is, agents prefer certain stimuli or states of the world to others. They seek out certain states or stimuli and avoid others.

There are intellectually demanding interpretations of these locutions which we will discuss later. However, as enactivists have pointed out (Di Paolo, Buhrmann, and Barandiaran 2017), all living organisms are goal-directed in a minimal sense. Even bacteria reliably avoid noxious stimuli, so can in a minimal sense be described as pursuing goals related to maintaining organismic integrity.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

The goal-directedness dimension captures this minimal sense of goal-pursuit. In this sense, a system is goal-directed when there are bodily or environmental states the system aims for in many different kinds of situation and that it seeks to maintain in the face of perturbation. The notions "aiming for" or "seeks to" should be interpreted in accordance with Dennett's (1991; 1987) account of propositional attitudes. According to this view, a system has certain goals when attributions of these goals are useful for predicting, understanding or controlling the system's behavior. No deeper cognitive state is needed to justify such attributions. [2]

According to this notion, all living organisms have some degree of goal-directedness. However, not all artefacts do. A table does not move, so it does not make sense to ascribe goals to it. A pocket calculator does calculations, but describing pocket calculators as having the goal to make correct calculations does not add any predictive or explanatory value, since there are no states they continuously seek out or maintain. Hence, we should not count them as goal-directed.

Yet, according to this notion, many current AI systems have goals. For instance, a chess-playing system can sometimes be better understood by ascribing goals like "capture the opponent's queen" to it (Dennett, 1987). We can sometimes interpret reinforcement learning systems fruitfully as having the goal to maximize rewards. Foundation language models of sufficient complexity can best be predicted by attributing the goal of accurate text prediction to them.

However, we can distinguish minimal from high-level goal-directedness. A high level of goal-directedness requires so-called goal-directed, as opposed to habitual, control of behavior. In goal-directed control, systems employ independent representations of the value of

[2] Similar to Dennett (1991), I don't hold that goal-directedness exists only relative to an interpreter. Instead, minimal goal-directedness consists in objective patterns of behavior which can be discerned from the intentional stance.

outcomes and of dependencies between actions and outcomes to calculate the expected value of possible actions (Butlin 2020). They then select the action with the highest expected value.

By contrast, in habitual control, values of actions themselves are learned and used in action selection.[3] The system learns which actions, in which situations, tend to cause valuable outcomes but does not explicitly represent the valuable states its actions might bring about. While habitual control still manifests preferences between world states, only in goal-directed control is the behavior of beings directly sensitive to the value of possible states of the world and thus goal-directed in a more demanding sense. It is not clear how widespread goal-directed control is, e.g. in the animal kingdom.

In conclusion, the dimensions account concurs with Dennettians that systems that are better predictable and explainable by attributing goals to them differ from others in ways constitutive of agency. At the same time, contra this interpretivist view, the dimensions account entails that agency is far from exhausted by being predictable and explainable in this way. Goal-directed control as well as the other agency dimensions exceed this minimal form of goal-directedness.

2.2.2 Autonomy

The second dimension is *autonomy*. A being is autonomous to the extent that it can function on its own, without intentional intervention by an outside observer. We can say that there are two levels of autonomy. A system is minimally autonomous if it can propel itself to behavior (Glock 2019). Mere tools, e.g. a TV, are inactive unless they are specifically and intentionally activated by someone else. Similarly, ChatGPT only starts acting once someone delivers a prompt to it. By contrast, animals don't need someone's explicit invitation to start being active. They can initiate their behavior themselves. The same goes for videogame AI systems which

[3] Computationally, this distinction corresponds to the distinction between particular forms of model-based and model-free reinforcement learning, respectively (Sutton and Barto 2018).

continuously behave in the virtual world without explicit prompting. Thus, a minimal form of autonomy seems to be common to all animals and is lacked by some, but not all, AI systems.

Moreover, a system has a high level of autonomy if, in addition, it relies more on learning about the world, and less on pre-specified, innate knowledge (Russell and Norvig 2020: 60). Thus, a certain degree of learning is required for a high degree of agency (Dretske 1999). For if many behavioral dispositions or internal representations are innate, then the behavior of the system is still controlled significantly by its designer, and not determined autonomously. If the system learns about and flexibly responds to its environment, then it can determine its behavior autonomously.

This does not mean that any degree of innate knowledge undermines high autonomy. The system only needs to have sufficient learning abilities and flexibility to be able to display behaviors and develop novel goals (in the minimal sense from last sub-section) which it has not been intentionally equipped with by its designer. Cats (or other mammals), which behave in non-stereotypical ways and gain new preferences over their lifetime, count as highly autonomous.

2.2.3 Efficacy

The third dimension is *efficacy* (Chan et al. 2023). A being is efficacious if its goal-directed behavior affects the world, without someone else's mediation or intervention. Efficacy encapsulates the idea that a being is more agentic if its behavior does not require support by others to matter. Animals are able to behave and change the physical world themselves. Similarly, it might be possible to develop robots which do not require human assistance to achieve their goals in the physical world. Animals and these robots would thus score high on efficacy in this sense.

However, many AI systems do not possess much, or any, efficacy. LLMs – at least when not augmented in some way – only give texts as outputs. Thus, they cannot intervene on the

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

world themselves. This is why they lack a minimal degree of efficacy. Second, videogame AI acts and exerts its influence on an environment, although it is only a virtual environment. Thus, it has low efficacy.

The efficacy of a system can be specified more precisely by the extent to which it requires another being, typically a human, to be in the control loop (Merat et al. 2019). Systems lack minimal efficacy if the human needs to be fully in the loop. In this case, the human exerts control over the physical or virtual environment whereas the AI system merely presents information to the human on which the latter then acts. Systems have low efficacy if the AI exerts control over the physical or virtual environment, while the human is "on the loop" (Merat et al. 2019), i.e. supervises, corrects and assists the AI behavior. Finally, systems are high in efficacy if they can directly and successfully act on the physical world to achieve their goals, without human assistance or monitoring.

The efficacy dimension does not imply that human paraplegics are diminished in agency, as they have many means to directly affect the world surrounding them. However, it does entail that humans with locked-in syndrome, who cannot move at all except for communicating with their eyes, have diminished agency. They are agents and have high values on all other dimensions but lack efficacy.

2.2.4 Planning

The fourth dimension is *planning*. Following Chan et al. (2023), I will understand the planning dimension as 'the degree to which the algorithmic system is [able] to make decisions that are temporally dependent upon one another to achieve a goal and/or make predictions over a long time horizon'. That is, the longer a system can plan in advance, the higher it scores on the planning dimension. This ability makes the system more agentic, as it increases its ability to act in accordance with its own long-term goals and to detach from the specific inputs the current situation confronts it with.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

As emphasized by Butlin (2022), planning requires a sensitivity to instrumental value. That is, a planning system must understand that achieving a certain outcome might be good for it, not (only) because it is intrinsically desired, but because it makes it possible to later attain a desired state. Otherwise, plans having multiple consecutive steps cannot get off the ground.

Following Butlin, I hold that behaving in accordance with instrumental relationships only counts as an instance of agency if the behavior is learned, not if is hard-wired. This is because planning only contributes to agency if it is associated with at least a minimal degree of flexibility. The relevant degree of flexibility requires that the system is able to *learn* instrumental relations which shape behavior, i.e. that representations of these instrumental relations are not all hardwired.

Hence, any system has a minimal degree of planning capacity if it has a learned sensitivity to instrumental value. As Butlin (2022; 2023) explains, reinforcement learning systems possess this sensitivity, because they are trained to perform behavior which leads to the greatest cumulative reward, not the greatest immediate reward. They learn to produce actions which do not lead to the maximum reward directly attainable, but are instrumentally useful for producing subsequent inputs enabling high reward later.

It is to be expected that such minimal planning is found in all mammals and a wide range of other animal species, too. Higher degrees of planning, requiring the capacity to produce coherent plans over long time scales, can be expected to correlate strongly with the general intelligence (Shevlin 2020) of animals and AI systems.

2.2.5 Intentionality

The fifth dimension is *intentionality*. It corresponds to the classical philosophical conception of intentional action (Schlosser 2019). One is an intentional agent in this sense if and only if one is able to act for reasons.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

Butlin (2023) argues that model-based reinforcement learning (RL) systems count as acting for reasons. The state transition function in a RL training environment specifies which states follow (with a particular probability) from which actions, given the states in which the actions have been produced. Model-based RL systems represent the state transition function along with the value of different states and use this information to choose which action to output. Consequently, Butlin claims, 'they represent and act on the basis of facts which count in favour of their actions, given their goals, and therefore […] act for reasons' (Butlin 2023).

However, the property Butlin refers to merely amounts to the capacity for goal-directed control which is – in the dimensions account – already included in the goal-directedness dimension. As Butlin himself grants, philosophers conceive of intentional action often in a more demanding way (e.g. Davidson 1963; Korsgaard 2008; 2018). In this tradition, human reasons for action are typically taken to be, or to be represented by, beliefs and desires, i.e. propositional attitudes. This is because reasons for action arguably need to be (based on) personal-level mental states which cause behavior. Otherwise, behavior driven by calculations of possible outcomes through sub-personal mental states, e.g. motor representations in the brain which the subject can neither reflect on nor consciously experience, might count as intentional action. [4]

We will assume here that reasons are, or are represented by, mental states which cause behavior and are owned by the system, i.e. the mental states need to be, or represent, reasons *for* the system. We can call these mental states *beliefs*. However, we should guard against the anthropocentric assumption that states need to exactly correspond to human propositional attitudes to count as beliefs in the relevant sense. For comparative purposes, I adopt the account of Newen and Starzak (2022) of what it takes for a cognitive system to have beliefs. In a nutshell, the system must have:

[4] On this, see also Jonathan Birch's talk on the evolution of reasons: https://drive.google.com/file/d/1Yi9DdHGHjdZyxVrt1Z_altuOzlNv0u1e/view (Accessed May 1, 2023).

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

> (a) informational states that are sufficiently decoupled from motivational states and that are (to varying degrees) intercombinable with other informational states as well as with different motivational states. These informational states have (b) a minimal structural organization, typically (but not exclusively) involving elements that represent objects, properties, or substances. Moreover, these states are (c) connected with minimal epistemic dispositions including (i) a sensitivity for new information; (ii) an ability to cluster or categorize new information when it concerns the same situation type or the same object, and so forth; and (iii) the ability to adjust the relevant informational states in the light of new evidence (Newen and Starzak 2022).

While this account is demanding, it is conceivable that some present-day or near-future model-based RL systems might possess the requisite ability to flexibly combine informational states and have all relevant epistemic dispositions.

However, merely having beliefs is not sufficient for these beliefs to constitute reasons *for the system*. In accordance with the traditional philosophical view, I take it that, in the human case, this often presupposes the ability to reflect on one's reasons, so as to reflectively endorse (Birch 2022) them, or to consciously call them to mind. If we want our conception of agency to not presuppose an understanding of consciousness (see section 2.3), acting for reasons might specifically require the abilities to represent one's reasons *as reasons* (Korsgaard 2008, pp. 3-5), to reflect on them and to approve of or reject them as a result. If so, a significant degree of metacognition would be necessary for agency.

In conjunction, my proposal is that the possession of beliefs and the ability to reflect on one's reasons as such and to revise or endorse them on this basis are jointly sufficient for being an intentional agent. This account assumes, controversially (Bourget and Mendelovici 2019), that intentional agency does not require phenomenal consciousness. Section 2.3 explains why it is theoretically advantageous to postulate that the capacity for intentional action does not presuppose consciousness. However, whether this assumption is in fact true is an open empirical, and metaphysical, question.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

Finally, it is an open question whether there are other ways in which a system can come to own its reasons and thus be an intentional agent without such sophisticated cognitive capacities, or whether the proposed account also specifies necessary conditions for intentional agency.

Dimensions of agency	Relevant features	Minimal examples
Goal-directedness	Having preferences and acting on them; goal-directed control	All living beings; chess engine
Autonomy	Ability to propel itself to behavior; learning about the world	All animals; game-playing reinforcement learning agent
Efficacy	Ability to affect the world, without someone else's mediation or intervention	All animals; goal-directed robots
Planning	Long-term planning; learned sensitivity to instrumental value	Intelligent animals, e.g. scrub jays; reinforcement learning agents
Intentionality	Acting for reasons; having propositional attitudes; metacognitive reflection	Humans; perhaps intelligent animals

Figure 2. This table depicts the five agency dimensions described in the text, features which determine how a system scores on the respective dimension and examples of beings which possess a minimal degree of agency on the respective dimension.

2.3 Relations between the five dimensions and to consciousness

Let us reflect on how the different dimensions hang together. It is notable that the first three dimensions all refer to features which are, at least in a minimal form, present in all living beings.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

Even bacteria prefer certain states to others and can move and have effects on the world without human intervention. By contrast, planning and intentionality (as characterized here) are features which depend, even in their minimal form, on a non-trivial degree of intelligence (Coelho Mollo 2022). This makes the first three dimensions less useful when characterizing animal agency. However, since not all artificial systems are goal-directed, autonomous and efficacious, these dimensions are fruitful when depicting different profiles of artificial agency.

It seems that goal-directedness is the most basic of these five dimensions. If a system is not goal-directed to a minimal extent, then it cannot pursue long-term goals and act for reasons, which would need to be connected to goals. Strictly speaking, the system might nevertheless behave autonomously and efficaciously, in the sense of not depending on outside intervention for its behavior and for affecting the outside world. An example would be a robot which selects its movements entirely randomly.

However, without goal-directedness, efficacious and autonomous behavior would have no systematic purpose. Real independence from outside control and intervention arguably requires having one's own goals which one pursues independently. For this reason, I hold that having a non-zero value on the other four dimensions presupposes a minimal degree of goal-directedness. Hence, if a system is not minimally goal-directed, it is not an agent.

By contrast, the other dimensions are conceptually independent from each other. Conceptually, for any proper subset of the other four dimensions, it is possible to score high on those dimensions and low on any of the others. Other dimensions further specify the abilities for pursuing goals that a system has: to pursue them autonomously, to affect the real world in their pursuit, to adhere to long-term goals and to own them as reasons.

A notable feature of the dimensions account is that it specifies agency independently of phenomenal consciousness, i.e. subjective experience (Nagel 1974). This independence is conceptual: none of the dimensions mentioned presupposes conceptually that goals, or any other states, are consciously experienced by the agent. This raises the intriguing possibility of

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

unconscious beings which score maximally on all agency dimensions (*full agency*). However, it is an empirical question whether unconscious full agency is nomologically possible, or whether certain forms of agency require conscious experience.

Characterizing agency independently of consciousness is theoretically desirable for two reasons. First, consciousness is notorious for being hard to grasp and to empirically investigate (Searle 1997). If assessments of agency would depend on assessments of consciousness, then it would be equally contestable what agency is, and which beings have it. Second, the independence from consciousness allows the concept of agency to play independent explanatory roles. It can provide an understanding of cognitive differences between kinds of animals and AI systems which deviates from the differences specified in debates on non-human consciousness, and it can potentially contribute to other debates in AI ethics (see section 4).

Why should we posit those five dimensions as the dimensions of agency and not accept a different theory? The answer has to lie in the usefulness and explanatory promise of the resulting account. To evaluate the merits of the dimensions account, I will first show that it has six virtues which other accounts of agency may not share. Then, I will suggest that the account has the potential to illuminate debates on the moral status of AI and AI existential risk.

3. Virtues of the dimensions account

3.1 Metaphysical and pragmatic approaches to agency

In support of the dimensions account, I argue that is has six important virtues. It is preferable to many alternatives because many accounts do not share these virtues. This project can be interpreted either metaphysically or purely pragmatically: Metaphysically, the aim of an account is to capture the nature of agency. The virtues would be understood epistemically: because the dimensions account has these virtues, it is more likely to be true of agency. On such a view, the usefulness of an account of agency is evidence for its truth, because the truth of an account would explain why it is useful.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

Pragmatically, the aim of the account is to provide a notion of agency which is useful for comparative research as well as the ethical-societal aims connected to agency, e.g. the ones outlined in section 4.[5] The virtues would be understood pragmatically: in part because the dimensions account has these virtues, it is scientifically and societally useful. The pragmatic approach denies that an account has to track independent facts about the nature of agency to be useful.

I will motivate the dimensions account based on its usefulness, including its explanatory promise. However, the account itself and the virtues used to motivate it are neutral between the competing metaphysical and pragmatic approaches to agency. It is open whether usefulness is regarded as an end in itself or should be seen as to be explained by the putative fact that the account captures independent facts about the nature of agency.

A distinctive task for the dimensions account is to contribute to a deeper understanding of *artificial* agency. Hence, since I aim for an account which not only encompasses human and animal agency but also agency in artificial systems, it needs to be more general than many proposals in the literature. To this end, I will assume that it is conceptually possible for artificial systems to be agents. This assumption is justified on pragmatic and explanatory grounds. If an account of agency rules out artificial agency completely, it cannot describe how artificial systems might differ in agency-relevant ways which curtails its usefulness and prevents it from explaining cognitive, behavioral, and ethical differences between AI systems.

In the remainder of the section, I will outline certain desirable qualities accounts of agency may have ("virtues") and then argue that the dimensions account possesses these virtues.

3.2 Origin-neutrality and multiple realizability

[5] The pragmatic approach can be understood as *conceptually engineering* agency (Cappelen 2018; Koch, Löhr, and Pinder 2023).

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

First, an account of agency should be, to some extent, *origin-neutral* (1). Agency could have different origins in different beings. Obviously, beings which are produced through biological evolution by natural selection can be agents. For humans are agents. However, an account of agency should be able to encompass beings which lack an evolutionary origin. Otherwise, it cannot accommodate artificial agency as well as other putative forms of agency, such as group agency (List and Pettit 2011; Pettit 2009).

Second, agency is, at least in principle, *multiply realizable* (2) (Polger & Shapiro, 2016, ch. 3). That is, agency could be sustained by different underlying structures (e.g. biological or silicon-based). Importantly, a proper investigation of agency might bring out that agency is actually not origin-neutral or not multiply realizable. However, this would be an empirical discovery. An account of artificial agency should not rule out *a priori* that AI systems can be agents.

That multiple realizability is a virtue is suggested by a focus on artificial agency. While the capacities enabling agency in humans are based on the biological brain, artificial agency would depend on other processes. In addition, developmental differences in humans suggest a more modest form of multiple realizability. The brains of human children differ from the brains of adults in some ways, but both can be agents. Finally, rejecting that multiple realizability is an in-principle possibility would implausibly entail, e.g. that AI systems behaviorally and functionally identical to humans are not agents.

The dimensions account clearly has these two virtues. All five dimensions characterize agency in an origin-neutral manner. Goal-directedness, autonomy etc. can, at least conceptually, all have their origin in evolution by natural selection, intentional human design or a variety of other sources. The same goes for virtue 2. All five dimensions pick out properties which are multiply realizable. It is perfectly possible to hold that goal-directedness, autonomy etc. are in some beings not sustained by biological brains, but instead by non-biological substrates (e.g. silicon).

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

3.3 Distinctive epistemic role

Third, the notion of agency should play a *distinctive epistemic role* (3), e.g. for scientific explanation and ethical understanding. This means that an account of agency should not describe agency as a property which is reducible to some other psychologically or philosophically important property like sentience (Wilcox 2020), cognition or rationality. Moreover, agency should have utility or explanatory power independent of its relation to these other properties. An account of agency would violate this constraint if it would entail that a system's being an agent is only philosophically or scientifically relevant because it makes it more likely that the system is, e.g., cognitive or rational. In other words, an account of agency has this virtue only if it characterizes a distinctive set of explananda or explanantia in scientific, philosophical, or everyday explanation and reasoning.

Likewise, the dimensions account has virtue 3 since the notion of agency which is captured by the five dimensions plays a distinctive epistemic role. That it, it is not reducible to other important concepts such as intelligence, sentience or cognition. The easiest way to illustrate this is to point out that agency – on the dimensions account – is not co-extensive with any of these other concepts. First, while all living beings count as minimal agents on my account, it seems that not all living beings are intelligent or sentient (e.g. bacteria). There are views on which all living beings are cognitive, however (Thompson 2010). Yet, the property *being cognitive* is not typically regarded as admitting degrees which distinguishes it from the notion of agency as construed here. In addition, there is no account of cognition I know of whose degrees of cognition map well onto the dimensions of agency described here. Furthermore, the dimensions account does not make the utility or explanatory power of agency dependent on its relation to some other important philosophical or scientific concept, such as cognition.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

3.4 Faithfulness to established usage

Fourth, the notion of agency should be *faithful to established usage* (4) in science, philosophy and folk discourse. If an account of agency deviates from typical usage of the notion of agency, this should be justified by proportional theoretical advantages. For instance, the account should have the virtues characterized here or play the ethical role outlined in section 4 to a higher degree. If an account is overly revisionary, it might lose sight of what is distinctive about agency which might compromise its epistemic value. It might fail to explain the phenomena, e.g. the distinction between merely behaving and being an agent, which are the original explananda of the notion of agency.

That being said, I set aside literature on AI *moral* agency and on the conditions under which humans ascribe agency to artificial systems. The former concerns a specific form of agency, acting on moral reasons, which is not required for agency in general. The latter line of research shows that human attributions of mental states to AI systems are significantly driven by superficial features (Küster, Swiderska, and Gunkel 2021; Nijssen et al. 2019; Thellman, de Graaf, and Ziemke 2022), e.g. physical appearance, which do not matter to a philosophical and scientific account of agency. As such, in cases of conflict, I tend to privilege the philosophical and scientific use of agency over folk usage of the term.

The dimensions account has virtue 4 because each dimension has been related to agency by previous research. Moreover, the dimensions account respects influential contemporary theories of agency. First, goal-directedness is typically posited as the key defining feature of agency (Ferrero 2022; Glock 2019; Liljeholm 2021; Wilcox 2020). Second, the view that minimal goal-directedness is shared by all living beings is prominent in enactivism (Di Paolo, Buhrmann, and Barandiaran 2017). Third, connections of autonomy, efficacy and planning to agency are also commonly drawn (Butlin 2022; Chan et al. 2023; Dretske 1999). Fourth, intentionality captures the classical notion of agency in the philosophical literature (Schlosser

2019). For all these reasons, the account presented is faithful to previous conceptions of agency. Thus, it qualifies as a distinctive view of agency, rather than some other property.

3.5 Determinacy

Fifth, the notion of agency should be *determinate* (5). It has to be sufficiently precise that there is, for most kinds of beings, a fact of the matter regarding whether they are agents. Moreover, it should be sufficiently specific that – in principle and ideally in practice – it is possible to come up with empirical indicators of agency in various systems. Ceteris paribus, the more amenable to empirical investigation a notion of agency is, the better. Of course, there may be and likely are borderline cases of agency. But the usefulness of an account of agency would be compromised if it would entail that, for *most* beings, it is indeterminate whether they are agents, or to what degree they are agents.

Some discussion is instructive to showcase that the dimensions account has the determinacy virtue. Previously, I have already given examples regarding the value different beings might have with respect to particular agency dimensions. For instance, I have claimed that all living beings possess minimal goal-directedness, that language models (without augmentation) lack autonomy and efficacy and that planning correlates with general intelligence. This suffices to demonstrate that the notion of agency developed here is not completely indeterminate: there are many definite attributions of agency that can be made.

Moreover, this account allows even more precise judgements and is fruitful for empirical investigation of the distribution of agency. This is because all agency dimensions are accessible to empirical study. Whether a system has minimal goal-directedness can be inferred from its overt behavior: If the system's behavior serves to repeatedly seek out certain states and stimuli, or avoid them, the system is minimally goal-directed. Whether a system exhibits goal-directed control, as opposed to habitual control, can be ascertained either by examining the algorithms it implements or its behavior. For the former, one can examine whether the system directly

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

learns values of possible actions or does something akin to calculating the expected value of actions, based on their potential outcomes. For the latter, goal-directed control is already studied empirically in animals, for instance via outcome devaluation experiments (Adams and Dickinson 1981; Butlin 2020).

Autonomy and efficacy are features typically[6] manifest in a system's overt behavior. By observing a system, one can tell to what extent it relies on the intentional assistance of others to initiate activity and to affect the world.

How a system scores on the planning dimension can be determined via knowledge about its algorithm, the training it has received or its behavioral capacity. As an example of the first method, consider that knowing that AlphaGo performs Monte Carlo simulation is sufficient to know that it has some planning abilities (Halina 2021). Similarly, knowing that a particular chatbot cannot transfer knowledge gained in one dialogue to other interactions makes it clear that its ability for planning over longer time-horizons is severely constrained. Second, typically, training via reinforcement learning ensures that the system learns sensitivity to instrumental relations, as the system is trained to maximize cumulative rewards over episodes of interaction with the environment (Butlin 2022). Third, we can judge whether the system consistently behaves in a way effective at achieving its goals over long time horizons, to the detriment of less important short-term goals. If so, it possesses long-term planning ability.

With respect to the intentionality dimension, relevant abilities are the possession of beliefs and metacognitive reflection. The conditions on belief possession by Newen and Starzak (2022) are straightforwardly cognitive and open to empirical research on AI systems (and also animals). Clues on a being's ability to reflect on its reasons can be gained based on a combination of the flexibility and intelligence of its decision-making and dedicated tests of non-human metacognition (Carruthers and Williams 2019). In the case of AI systems, explicit verbal

[6] Cases of AI deception (Park et al. 2023), particularly from conceivable future AI which possesses theory of mind abilities, would be an exception.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

reasoning abilities and methods to track the system's internal representations (e.g. Zou et al. 2023) can furthermore be used. Hence, a solid grasp of the behavioral capacities of an AI, perhaps in conjunction with the algorithms governing its behavior, may be sufficient to determine how it scores on the intentionality dimension.

3.6 Informativeness

Finally, an account of agency should be *informative* (6). The kind of informativeness which especially interests us here concerns the standpoint of comparative cognition research. If an account of agency is informative, then it can convey much information about the agency-relevant properties and capacities of humans, animals and AI systems which are, from a comparative standpoint, interesting and important. In other words: An informative account of agency allows characterizations of human, animal and AI agency which contain much relevant information. It can describe, with sufficient level of detail, not only which beings are agents, but how the kind of agency we find in different beings differs.

Two types of accounts are relatively uninformative, in this sense. First, these are accounts of agency which combine two features: First, either only humans or (virtually) all animals and AI systems possess agency, according to the account. Second, agency does not admit of (scientifically interesting) degrees. Such an account has very limited descriptive and explanatory resources: it cannot describe fine-grained differences in agency between humans and various kinds of animals and AI systems. For this reason, it can only explain either differences between humans on the one hand and all animals and AI systems on the other, or no differences between humans, animals and AI systems at all. By contrast, a more informative account can explain differences between humans and different kinds of animals and AI systems in virtue of differences in the possession of agency and their kinds and degrees of agency.

A different, related way in which an account could be uninformative is by being too broad: For example, as we saw, everyone agrees that agency is connected to the capacity to

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

pursue goals. However, without elucidating what specifically this capacity amounts to, an account of agency in terms of the capacity for goal pursuit is uninformative.

Considering informativeness a virtue can be justified pragmatically as well as metaphysically. Pragmatically, an informative account allows to give more detailed characterizations of agency in different systems and to draw more fine-grained distinctions. This will make, all other things being equal, an informative account of agency more useful. Metaphysically, an account of agency should provide an accurate and detailed characterization of what agency really is. If agency is a complex and multi-facetted property, then only an informative account can satisfactorily describe it. Finally, as we have seen, informativeness is connected to explanatory power. An informative account has more resources to explain phenomena in recourse to agency.

A distinctive feature of the dimensions account is that it allows for fine-grained and nuanced characterizations of agency which makes it very informative. Instead of delivering only yes-no judgements regarding agency in animals or AI, the account supports degreed and multi-facetted assessments. It makes room for specific judgements of the form:

(I) System A has higher agency on dimension 1, but system B has higher agency on dimension 2.

That is, attributions of agency can vary according to the relevant degree of agency as well as the relevant dimension. An assessment of agency which tells us which value a system has on each agency dimension provides knowledge of the overall agency profile of a system. Consequently, it is much more informative than an account which merely allows for the judgements ‘X is an agent’ and ‘X is not an agent’.

In addition, the dimensions account can also make sense of such flat, undifferentiated statements. Take this statement:

(II) System A has higher agency than system B.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

According to the dimensions account, this statement can be interpreted in two ways. First, it might implicitly refer to a particular agency dimension. In this case, we need to make this tacit reference explicit. However, the statement (II) – taken literally – also has definite truth conditions: If system A has at least equal agency on all dimensions and higher agency on some dimensions than system B, then (II) is true.

Finally, consider:

(III) System A is not an agent.

According to the dimensions account, this statement is true if A fails the criterion for minimal agency on each dimension. This may require testing for each dimension separately. However, as I have argued, minimal goal-directedness may be seen as a necessary condition for a non-zero value on each of the other dimensions. If so, (III) is true if and only if system A is not minimally goal-directed.

Thus, the dimensions account of agency is more informative than alternative non-dimensional accounts. It possesses the resources to evaluate flat, undifferentiated claims entailed by non-dimensional accounts, but in addition allows for more fine-grained assessments of a system's agency profile, where necessary.

In conclusion, the dimensions account has six important virtues, which are depicted in Figure 1. After having shown that the dimensions account has these virtues, I will now argue that it may help illuminate two important issues in AI ethics.

Nr.	Virtue
1	Origin-neutrality
2	Multiple realizability
3	Epistemic distinctiveness
4	Faithfulness (to established usage)

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

5	Determinacy
6	Informativeness

Figure 1. This table summarizes the six virtues of the dimensions account characterized in the main text.

4. AI agency and AI ethics

Agency is not only of scientific interest. Recent research indicates that agency, as characterized by the dimensions account, may be suited to play an important role in AI ethics. This concerns particularly two questions: First, there is a recent debate on robot rights and the moral status of AI systems (Gunkel 2018; Ladak 2023; Moosavi 2023; Müller 2021; Schwitzgebel and Garza 2015; Shevlin 2021). A being possesses moral status if and only if it matters morally for its own sake (Jaworska and Tannenbaum 2021). That is, if a being has moral status, we have obligations to that being in virtue of its intrinsic properties, not because it matters to someone else.[7]

According to the dominant view, moral status is determined by sentience, i.e. the capacity to undergo pleasant and unpleasant conscious experiences such as joy, pain and fear (Kriegel 2019; Nussbaum 2007; Singer 2011). However, there also has arisen vocal opposition to the view that sentience plays this normative role (Danaher 2020; Gunkel 2019; Kammerer 2019; 2022; Shepherd 2023).[8] In particular, while most agree that sentience is sufficient for moral status, there is less agreement that sentience is necessary (Kagan 2019; Roelofs 2022).

Various authors have recently expressed sympathy for the view that some form of agency, construed as being independent of consciousness, suffices for moral status (Birch 2022; Delon 2023; Kagan 2019; Kammerer 2022; Ladak 2023). I do not commit to this view here. Nevertheless, the dimensions account contributes to this discussion because it supplies a notion of agency which can be used to make sense of this claim. It allows fruitful discussion of the

[7] Moral status, as defined here, is synonymous with moral patiency. It is logically independent of moral agency, i.e. the capacity to act for moral reasons.
[8] For a defense of the view that sentience is normatively significant, see Dung (2022; forthcoming).

question which degree of agency on which dimension, if any, may suffice for (which degree of) moral status. For instance, minimal goal-directedness is a necessary condition for having interests which are grounded in agency. Therefore, if a system does not prefer and avoid certain things, we cannot say that it has moral status, or at least no moral status grounded in agency. Minimal goal-directedness is not sufficient for moral status, however, because it does not entail that a being's interests matter to it in a morally relevant sense (Ladak 2023).

Planning may also be relevant to moral status. If a system scores high on the planning dimension, it can have and pursue long-term goals. Thus, it unlocks a new class of interests, relevant to a system's long-term prospects. These interests might be especially relevant for the right of a system to not be killed. Arguably, one thing that makes death bad is that it leads to the frustration of long-term goals (Luper 2021; Singer 2011). However, again, whether interests concern the more distant future does not affect whether they matter morally. So, systems with high planning abilities may lack moral status.

The intentionality dimension might provide an answer to the question which beings have interests that matter morally. For instance, one may hold that a being's interests are morally relevant if it is capable of *reflectively endorsing* these interests (Birch 2022). By reflecting on its interests, and deciding to endorse them, a being adopts an attitude of care towards its interests. For this reason, its interests might start to matter morally. The capacity to reflect on one's own mental states and to rationally endorse or reject them is what the highest level of intentionality consists in. In this way, the dimensions account can clarify the claim that beings can have moral status, or a particular degree of moral status, in virtue of agency. It tells us which aspect of agency would need to be sufficient for moral status for this claim to be true.

Second, the account of agency developed here may be relevant to debates about catastrophic and existential risk from AI. A significant number of authors has expressed concern that advanced AI poses an existential risk to humanity (e.g. Bostrom 2014; Carlsmith 2022; Dung 2023a; Dung 2023b; Ngo, Chan, and Mindermann 2022; Russell 2019). In a

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

nutshell, the typical worry is that we might build AI which is intellectually superior to humans and optimizes for goals which do not explicitly involve human flourishing. If so, it may start to exhibit power-seeking behavior, e.g. improving itself and accumulating resources, to be more effective at optimizing for its goal (whatever that may be). If successful, so the thought, it will then disempower humanity.

However, intuitively, current state-of-the-art models don't seem like they raise real-world threats of massive power-seeking.[9] Again, looking at LLMs is most instructive. While LLMs have striking cognitive capacities, they are – in a sense which needs to be made more precise – mere tools. They don't seem to pursue goals in the way humans or other animals do. A natural suggestion is that we should be worried about existential risk from power-seeking AI if AI systems become sufficiently intelligent *and* develop agency.

I will make the controversial assumption that this suggestion is on the right track. If so, the dimensions account of agency can potentially illuminate which particular kind of system we should be concerned about. It appears that at least four of the five agency dimensions are relevant to the degree to which an AI system poses an existential risk. First, minimal goal-directedness can be regarded as a necessary condition for power-seeking. If behavior does not aim at goals at all, then it will not systematically aim to accumulate power. Next, autonomy and efficacy are plausibly important factors contributing to dangerous forms of power-seeking. Let us take autonomy first. If a system cannot initiate activity itself or does not learn from experience, it cannot – for a longer time period – behave in ways which are at odds with the interests of its designers and users, if they realize that the behavior is not in their interest. Without autonomy, a system does not have the degree of independence necessary to maintain an adversarial relation to humanity. In particular, if the system depends on human prompting,

[9] That being said, text outputs describing power-seeking intentions have been shown to spontaneously emerge in LLMs in some contexts (Perez et al. 2022).

a scenario where it independently continues to amass resources to the detriment of humanity is ruled out.

Similarly, if the system is not efficacious, i.e. relies on humans to affect the world, there again always is a human in the loop. Hence, if humans notice that an AI system begins to act against their interest, they can stop assisting the AI system and consequently limit its real-world influence. Thus, when efficacy is low, scenarios where an AI – removed from human oversight – accumulates resources and ultimately disempowers humanity are apparently ruled out.

While autonomy and efficacy are thus important factors to consider regarding existential risk from AI, the absence of these properties does not rule out AI existential risk (if there is such a risk in the first place). This absence is compatible with scenarios in which AI systems are intentionally used by some humans to suppress or kill others or where AI systems find ways to cause massive harm by deceiving the humans they interact with. Even more clearly, non-efficacious and non-autonomous systems may furthermore be dangerous because they might, with time, find ways to become more autonomous and efficacious.

In addition, the planning dimension is important for understanding existential risk from AI. To be able to disempower humanity, an AI would likely need to be able to form and execute long-term plans. To the extent that the system cannot execute on long-term plans, there is low risk that it can accumulate power to achieve some other goal. For this process would most likely involve planning a large number of steps into the future, and then executing this plan carefully.

The relevance of the intentionality dimension is unclear. That being said, one may argue that systems able to reflect on their goals and to revise them are less predictable and controllable and therefore more dangerous. On the other hand, it has been argued that reflective AI systems may be less susceptible to dangerous forms of power-seeking (Müller and Cannon 2022). In both directions, the relevance of the intentionality dimension to AI existential risk would need to be explored systematically. Thus, all dimensions of agency may comprise features which contribute to risks of AI power-seeking.

Preprint. Please cite the published version: https://doi.org/10.1093/pq/pqae010

All the main claims discussed in this section are controversial: agency may be irrelevant to moral status, AI may not pose an existential risk and the existential risk posed by AI may be independent of agency. I cannot defend these claims here. However, the dimensions account supplies a notion of agency which can be used to make these claims precise and to discuss their merits. If it turns out that agency is relevant to AI moral status and existential risk, the dimensions account can be used in future research to make more fine-grained assessments of AI moral status and existential risk from AI, based on the notion of agency.

5. Conclusion

I this paper, I have developed a multi-dimensional account of artificial agency. This enables fine-grained comparative assessments of agency where a system's agency profile can be characterized as a point in a multi-dimensional space. The five dimensions are goal-directedness, autonomy, efficacy, planning and intentionality. I argued that this account is justified in virtue of its usefulness, particularly its explanatory power. The account shows how agency can be origin-neutral, multiply realizable and determinate and retains agency as a notion which is faithful to common usage, epistemically distinctive as well as informative. More speculatively, the account enables the notion of agency to contribute to questions regarding AI moral status and existential risk.

The last two questions are obviously ethically important. While there is a lot of uncertainty, creating AI systems which are – or might become – agents might be dangerous. Both because those systems might have moral status, so that we might harm them, and because they might exhibit power-seeking tendencies, so that they might harm us. At the same time, there are strong financial and scientific incentives to build systems which are increasingly agentic (Chan et al. 2023). This is worrisome. Nevertheless, it is important to not jump to conclusions about the appropriate ways for societies to confront these risks. A sustained, critical and rigorous debate is necessary.

References

Adams, C. D., and Anthony D. (1981) 'Instrumental Responding Following Reinforcer Devaluation', *The Quarterly Journal of Experimental Psychology Section B,* 33/2b: 109–21. https://doi.org/10.1080/14640748108400816.

Birch, J. (2022) 'Materialism and the Moral Status of Animals', *The Philosophical Quarterly*, January. https://doi.org/10.1093/pq/pqab072.

Birch, J., A. K. Schnell, and N. S. Clayton (2020) 'Dimensions of Animal Consciousness', *Trends in Cognitive Sciences,* 24/10: 789–801. https://doi.org/10.1016/j.tics.2020.07.007.

Bostrom, N. (2014) *Superintelligence. Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bourget, D., and A. Mendelovici (2019) 'Phenomenal Intentionality', in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, Fall 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2019/entries/phenomenal-intentionality/.

Butlin, P. (2020) 'Affective Experience and Evidence for Animal Consciousness', *Philosophical Topics,* 48/1: 109–27. https://doi.org/10.5840/philtopics20204816.

Butlin, P. (2022) 'Machine Learning, Functions and Goals', *Croatian Journal of Philosophy,* 22/66: 351–70. https://doi.org/10.52685/cjp.22.66.5.

Butlin, P. (2023) 'Reinforcement Learning and Artificial Agency', *Mind & Language*, May, mila.12458. https://doi.org/10.1111/mila.12458.

Cappelen, H. (2018) *Fixing Language*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198814719.001.0001.

Carlsmith, J. (2022) 'Is Power-Seeking AI an Existential Risk?', arXiv. https://doi.org/10.48550/arXiv.2206.13353.

Carruthers, P., and D. M. Williams (2019) 'Comparative Metacognition', *Animal Behavior and Cognition,* 6/4: 278–88. https://doi.org/10.26451/abc.06.04.08.2019.

Chan, A., R. Salganik, A. Markelius, C. Pang, N. Rajkumar, D. Krasheninnikov, L. Langosco, et al. (2023) 'Harms from Increasingly Agentic Algorithmic Systems', arXiv. https://doi.org/10.48550/arXiv.2302.10329.

Coelho Mollo, D. (2022) 'Intelligent Behaviour', *Erkenntnis*, May, 1–18. https://doi.org/10.1007/s10670-022-00552-8.

Danaher, J. (2020) 'Welcoming Robots into the Moral Circle: A Defence of Ethical Behaviourism', *Science and Engineering Ethics,* 26/4: 2023–49. https://doi.org/10.1007/s11948-019-00119-x.

Davidson, D. (1963) 'Actions, Reasons, and Causes', *The Journal of Philosophy,* 60/23: 685–700. https://doi.org/10.2307/2023177.

Delon, N. (2023) 'Agential Value', Substack newsletter. *Running Ideas* (blog). January 12, 2023. https://nicolasdelon.substack.com/p/agential-value?utm_campaign=auto_share.

Dennett, D. C. (1987) *The Intentional Stance*. Cambridge: MIT Press.

Dennett, D. C. (1991) 'Real Patterns', *Journal of Philosophy,* 88/1: 27–51.

Di Paolo, E., T. Buhrmann, and X. Barandiaran (2017) *Sensorimotor Life*, Vol. 1. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198786849.001.0001.

Dretske, F. I. (1999) 'Machines, Plants and Animals: The Origins of Agency', *Erkenntnis,* 51/1: 523–35. https://doi.org/10.1023/A:1005541307925.

Dung, L. (2022) 'Why the Epistemic Objection Against Using Sentience as Criterion of Moral Status Is Flawed', *Science and Engineering Ethics,* 28/6: 51. https://doi.org/10.1007/s11948-022-00408-y.

Dung, L. (2023a) 'Current Cases of AI Misalignment and Their Implications for Future Risks', *Synthese,* 202/5: 138. https://doi.org/10.1007/s11229-023-04367-0.

Dung, L. (2023b) *The argument for near-term human disempowerment through AI.*

https://philpapers.org/rec/DUNTAF-3
Dung, L. (forthcoming) 'Preserving the Normative Significance of Sentience', *Journal of Consciousness Studies*.
Dung, L., and A. Newen (2023) 'Profiles of Animal Consciousness: A Species-Sensitive, Two-Tier Account to Quality and Distribution', *Cognition,* 235 (June): 105409. https://doi.org/10.1016/j.cognition.2023.105409.
Ferrero, L., ed. (2022) *The Routledge Handbook of Philosophy of Agency*. New York: Routledge.
Glock, H. (2019) 'Agency, Intelligence and Reasons in Animals', *Philosophy,* 94/4: 645–71. https://doi.org/10.1017/S0031819119000275.
Gunkel, D. J. (2018) *Robot Rights*. Cambridge: The MIT Press. https://doi.org/10.7551/mitpress/11444.001.0001.
Gunkel, D. J. (2019) 'No Brainer: Why Consciousness Is Neither a Necessary nor Sufficient Condition for AI Ethics', in *AAAI Spring Symposium: Towards Conscious AI Systems*.
Hagendorff, T., L. N. Bossert, Y. Fai Tse, and P. Singer (2022) 'Speciesist Bias in AI: How AI Applications Perpetuate Discrimination and Unfair Outcomes against Animals', *AI and Ethics*, August. https://doi.org/10.1007/s43681-022-00199-9.
Halina, M. (2021) 'Insightful Artificial Intelligence.' *Mind & Language,* 36(2): 315–29. https://doi.org/10.1111/mila.12321.
Jaworska, A., and J. Tannenbaum (2021) 'The Grounds of Moral Status', in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, Spring 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/spr2021/entries/grounds-moral-status/.
Kagan, S. (2019) *How to Count Animals, More or Less*. Uehiro Series in Practical Ethics. Oxford: Oxford University Press.
Kammerer, F. (2019) 'The Normative Challenge for Illusionist Views of Consciousness', *Ergo, an Open Access Journal of Philosophy,* 6. https://doi.org/10.3998/ergo.12405314.0006.032.
Kammerer, F. (2022) 'Ethics Without Sentience. Facing Up to the Probable Insignificance of Phenomenal Consciousness', *Journal of Consciousness Studies*.
Kenton, Z., R. Kumar, S. Farquhar, J. Richens, M. MacDermott, and T. Everitt (2022) 'Discovering Agents', arXiv. https://doi.org/10.48550/arXiv.2208.08345.
Koch, S., G. Löhr, and M. Pinder (2023) 'Recent Work in the Theory of Conceptual Engineering', *Analysis*, August, anad032. https://doi.org/10.1093/analys/anad032.
Korsgaard, C. M. (2008) *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*. Oxford: Oxford University Press.
Korsgaard, C. M. (2018) *Fellow Creatures*. Vol. 1. Oxford University Press. https://doi.org/10.1093/oso/9780198753858.001.0001.
Kriegel, U. (2019) 'The Value of Consciousness', *Analysis,* 79/3: 503–20. https://doi.org/10.1093/analys/anz045.
Küster, D., A. Swiderska, and D. Gunkel (2021) 'I Saw It on YouTube! How Online Videos Shape Perceptions of Mind, Morality, and Fears about Robots', *New Media & Society* 23/11: 3312–31. https://doi.org/10.1177/1461444820954199.
Ladak, A. (2023) 'What Would Qualify an Artificial Intelligence for Moral Standing?', *AI and Ethics*, January. https://doi.org/10.1007/s43681-023-00260-1.
Liljeholm, M. (2021) 'Agency and Goal-Directed Choice', *Current Opinion in Behavioral Sciences*, Value based decision-making, 41 (October): 78–84. https://doi.org/10.1016/j.cobeha.2021.04.004.
List, C., and P. Pettit (2011) *Group Agency*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199591565.001.0001.

Luper, S. (2021) 'Death.' In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2021. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/fall2021/entries/death/.

Merat, N., B. Seppelt, T. Louw, J. Engström, J. D. Lee, E. Johansson, C. A. Green, et al. (2019) 'The 'Out-of-the-Loop' Concept in Automated Driving: Proposed Definition, Measures and Implications', *Cognition, Technology & Work,* 21/1: 87–98. https://doi.org/10.1007/s10111-018-0525-8.

Moosavi, P. (2023) 'Will Intelligent Machines Become Moral Patients?' *Philosophy and Phenomenological Research*, September, phpr.13019. https://doi.org/10.1111/phpr.13019.

Müller, V. C. (2021) 'Is It Time for Robot Rights? Moral Status in Artificial Entities', *Ethics and Information Technology,* 23/4: 579–87. https://doi.org/10.1007/s10676-021-09596-w.

Müller, V. C., and Michael Cannon (2022) 'Existential Risk from AI and Orthogonality: Can We Have It Both Ways?', *Ratio,* 35/1: 25–36. https://doi.org/10.1111/rati.12320.

Nagel, T. (1974) 'What Is It Like to Be a Bat?', *Philosophical Review,* 83/4: 435–50. https://doi.org/10.2307/2183914.

Newen, A., M. Griem, and S. Pika (2022) 'A Conceptual Framework for Empathy in Humans and Nonhuman Animals', in *Wittgenstein and Beyond*, by C. C. Pfisterer, N. Rathgeb, and E. Schmidt, 1st ed., 203–24. New York: Routledge. https://doi.org/10.4324/9781003202929-15.

Newen, A., and T. Starzak (2022) 'How to Ascribe Beliefs to Animals', *Mind & Language,* 37/1: 3–21. https://doi.org/10.1111/mila.12302.

Ngo, R., L. Chan, and S. Mindermann (2022) 'The Alignment Problem from a Deep Learning Perspective', arXiv. http://arxiv.org/abs/2209.00626.

Nijssen, S. R. R., B. C. N. Müller, R. B. van Baaren, and M. Paulus (2019) 'Saving the Robot or the Human? Robots Who Feel Deserve Moral Care', *Social Cognition,* 37/1: 41-S2. https://doi.org/10.1521/soco.2019.37.1.41.

Nussbaum, M. C. (2007) *Frontiers of Justice: Disability, Nationality, Species Membership*. Harvard: Harvard University Press.

Orseau, L., S. McGregor McGill, and S. Legg (2018) 'Agents and Devices: A Relative Definition of Agency', arXiv. https://doi.org/10.48550/arXiv.1805.12387.

Park, P. S., S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks (2023) 'AI Deception: A Survey of Examples, Risks, and Potential Solutions', arXiv. https://doi.org/10.48550/arXiv.2308.14752.

Perez, E., S. Ringer, K. Lukošiūtė, K. Nguyen, E. Chen, S. Heiner, C. Pettit, et al. (2022) 'Discovering Language Model Behaviors with Model-Written Evaluations', arXiv. https://doi.org/10.48550/arXiv.2212.09251.

Pettit, P. (2009) 'The Reality of Group Agents' in *Philosophy of the Social Sciences*, edited by C. Mantzavinos, 1st ed., 67–91. Cambridge University Press. https://doi.org/10.1017/CBO9780511812880.007.

Roelofs, L. (2022) 'Sentientism, Motivation, and Philosophical Vulcans', *Pacific Philosophical Quarterly* n/a (n/a). https://doi.org/10.1111/papq.12420.

Russell, S. (2019) *Human Compatible: Artificial Intelligence and the Problem of Control*. Viking.

Russell, S., and P. Norvig (2020) *Artificial Intelligence: A Modern Approach*. 4th ed. Pearson.

Schlosser, M. (2019) 'Agency' in *The Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta, Winter 2019. Metaphysics Research Lab, Stanford University. https://plato.stanford.edu/archives/win2019/entries/agency/.

Schwitzgebel, E., and M. Garza (2015) 'A Defense of the Rights of Artificial Intelligences', *Midwest Studies in Philosophy,* 39/1: 98–119. https://doi.org/10.1111/misp.12032.

Searle, J. (1997) *The Mystery of Consciousness*. New York Review.

Shepherd, J. (2023) 'Non-Human Moral Status: Problems with Phenomenal Consciousness', *AJOB Neuroscience,* 14/2: 148–57. https://doi.org/10.1080/21507740.2022.2148770.

Shevlin, H. (2020) 'General Intelligence: An Ecumenical Heuristic for Artificial Consciousness Research?', *Journal of Artificial Intelligence and Consciousness*, May. https://doi.org/10.17863/CAM.52059.

Shevlin, H. (2021) 'How Could We Know When a Robot Was a Moral Patient?', *Cambridge Quarterly of Healthcare Ethics,* 30/3: 459–71. https://doi.org/10.1017/S0963180120001012.

Singer, P. (2011) *Practical Ethics*. 3rd ed. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511975950.

Starzak, T. B., and R. D. Gray (2021) 'Towards Ending the Animal Cognition War: A Three-Dimensional Model of Causal Cognition', *Biology & Philosophy,* 36/2: 9. https://doi.org/10.1007/s10539-021-09779-1.

Sutton, R. S., and A. G. Barto. (2018) *Reinforcement Learning. An Introduction.* 2nd ed. Cambridge, MA, USA: MIT Press.

Thellman, S., M. de Graaf, and T. Ziemke (2022) 'Mental State Attribution to Robots: A Systematic Review of Conceptions, Methods, and Findings', *ACM Transactions on Human-Robot Interaction,* 11/4: 1-41. https://doi.org/10.1145/3526112.

Thompson, E. (2010) *Mind in Life. Biology, Phenomenology, and the Sciences of Mind*. Harvard: Harvard University Press.

Wilcox, M. G. (2020) 'Animals and the Agency Account of Moral Status', *Philosophical Studies,* 177/7: 1879–99. https://doi.org/10.1007/s11098-019-01289-x.

Zou, A, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, et al. (2023) 'Representation Engineering: A Top-Down Approach to AI Transparency', arXiv. https://doi.org/10.48550/arXiv.2310.01405.

Centre for Philosophy and AI Research, University Erlangen-Nürnberg, Germany