THE UNIVERSITY of EDINBURGH

Edinburgh Research Explorer

Intelligent capacities in artificial systems

Citation for published version:
Kasirzadeh, A & McGeer, V 2024, Intelligent capacities in artificial systems. in WA Bauer & A Marmodoro (eds), *Artificial Dispositions: Investigating Ethical and Metaphysical Issues.* 1 edn, Bloomsbury Publishing, pp. 141–168. https://doi.org/10.5040/9781350336148.ch-007

Digital Object Identifier (DOI):
10.5040/9781350336148.ch-007

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Peer reviewed version

Published In:
Artificial Dispositions

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 04. Jul. 2024

Intelligent capacities in artificial systems

Forthcoming in Artificial Dispositions: Investigating Ethical and Metaphysical Issues (editors: William A. Bauer and Anna Marmodoro), Bloomsbury Publishing.

Atoosa Kasirzadeh, University of Edinburgh/The Alan Turing Institute
atoosa.kasirzadeh@ed.ac.uk

Victoria McGeer, Australian National University/ Princeton University
vmcgeer@princeton.edu

Abstract. This paper investigates the nature of dispositional properties in the context of artificial intelligence systems. We start by examining the distinctive features of natural dispositions according to criteria introduced by McGeer (2018) for distinguishing between object-centered dispositions (i.e., properties like 'fragility') and agent-based abilities, including both 'habits' and 'skills' (a.k.a. 'intelligent capacities', Ryle 1949). We then explore to what extent the distinction applies to artificial dispositions in the context of two very different kinds of artificial systems, one based on rule-based classical logic and the other on reinforcement learning. Here we defend three substantive claims. First, we argue that artificial systems are not equal in the kinds of dispositional properties they instantiate. In particular, we show that logical systems instantiate merely object-centered dispositions whereas reinforcement learning systems allow for the instantiation of agent-based abilities. Second, we explore the similarities and differences between the agent-centered abilities of artificial systems and those of humans, especially as relates to the important distinction made in the human case between habits and skills/intelligent capacities. The upshot is that the agent-centered abilities of truly intelligent artificial systems are distinctive enough to constitute a third type of agent-based ability — blended agent-based ability — raising substantial questions as to how we understand the nature of their agency. Third, we explore one aspect of this problem, focussing on whether systems of this type are properly considered 'responsible agents', at least in some contexts and for some purposes. The ramifications of our analysis will turn out to be directly relevant to various ethical concerns of artificial intelligence.

Keywords. Dispositional properties; Artificial Intelligence; Object-centered dispositions; Agent-centered abilities; Skills; Habits; Artificial dispositions; Reinforcement learning systems; Responsibility; Responsibility gaps

1. Introduction

In this paper, we explore the nature of dispositional properties in the context of artificial intelligence (AI) systems. Following Gilbert Ryle (1949), we argue there is nothing more to 'being intelligent' than possessing a rich array of dispositional properties. But dispositional properties are not all of a piece. Intelligent creatures, notably human beings, possess dispositional properties of a characteristically complex kind that Ryle called 'intelligent capacities' or 'skills'. Our question, then, is to what extent such properties could be instantiated in artificial systems; hence, to be or become 'intelligent' in the manner of human beings. It is a subsidiary question as to whether artificial systems could be or become 'intelligent' in a different sense, however that comes to be defined.

At first glance, these may seem to be merely definitional matters. But that would be misleading. As we shall argue, possessing intelligent capacities in the human sense is a normatively substantive condition, underwriting the attribution of responsible agency. So to explore this question in relation to artificial systems is to explore the extent to which they can and should be considered responsible agents. This latter question lies at the heart of many ethical concerns raised by the promise (or threat) of our growing reliance on 'AI' systems in the modern world. Hence, in this paper we shall use the term "intelligence" or "intelligent" in a more restricted way than is implied by the acronym 'AI'; but for ease of reference, we shall also continue to use that acronym to refer more broadly to the variety of smart computational systems we discuss in this paper.

Our discussion will proceed as follows. In Section 2, we briefly sketch our background view of the general nature of dispositional properties, leaving many nuanced issues about them to one side. In section 3, we turn to Ryle's categorization of such properties, focussing on a fundamental distinction he makes between object-centered dispositions (i.e., properties like 'fragility') and agent-based abilities, with abilities encompassing both 'habits' and 'skills'. We discuss what is critical in these distinctions, focussing especially on why Ryle reserved the term 'intelligent' for skills alone. In the remainder of the paper (Sections 4-6), we explore the extent to which these Rylean distinctions apply to the properties of different kinds of artificial systems. In Section 4, we argue that artificial systems are not equal in the kinds of dispositional properties they instantiate. In particular, we show that logical systems (Nilsson, 1991) instantiate merely object-centered dispositions whereas reinforcement learning systems (Sutton & Barto, 2018) allow for the instantiation of agent-based abilities. In Section 5, we explore how the agent-centered abilities of reinforcement learning systems are like and unlike human abilities, raising substantial questions as to how we understand the nature of their

agency. In a final Section 6, we explore one aspect of this problem – namely, to what extent it makes sense to view such systems as 'responsible agents', at least in some contexts and for some purposes. We consider some of the ethical ramifications to answering this question in the way our analysis suggests.

2. The nature of dispositional properties

Dispositional properties are modal properties: they are properties of things that are attributable in virtue of how those things can be expected to behave under a range of characteristic conditions, actual or non-actual. Some examples: A vase is 'fragile' just in case there are a range of characteristic conditions under which it can be expected to break (strikings, throwings, droppings, etc.). A person is 'contagious' just in case there are a range of characteristic conditions under which they can be expected to infect others with a disease they are carrying. An animal is 'well-trained' just in case there are a range of characteristic conditions under which it can be expected to act in a way it was trained to do (e.g., sitting on command).

Notice that the attribution of modal properties seemingly has nothing to do with how frequently, if at all, the behaviour in question is manifested. A fragile vase may never break despite its fragility. By contrast, a well-trained animal may do what it was trained to do on a recurring basis. The difference between these two cases has nothing to do with what might be called the 'strength' of the dispositional property – how likely it is to manifest under the relevant range of characteristic eliciting conditions. Rather it has to do with how frequently these eliciting conditions obtain in the actual world. Hence, though a well-trained animal may often display the trained behaviour (e.g., sitting on command), it may not be as strongly disposed to sit on command as the fragile vase is disposed to break. That is to say, the probability of the animal's sitting is lower than the probability of the vase's breaking under conditions that are characteristic for each. Theorists may capture this idea by speaking of the degree to which the relevant dispositional property is 'sure-fire' under the relevant eliciting conditions.

There is another degree of variation for dispositional properties beyond their 'strength'– we might call this their 'robustness'. Robustness concerns the range of conditions under which the thing in question would manifest the relevant behaviour: the wider the range of such conditions, the more robust the dispositional property. For instance, a dog may be strongly disposed to sit on command, manifesting this behaviour quite reliably under certain specific conditions; but it may not be robustly disposed to sit on command, as it does so only in response

to its master's voice. Other people may tell it to sit, but that command falls, as it were, on deaf ears. The dog is sensitive only to its master voice, thereby manifesting this dispositional property to sit on command less robustly than other dogs that sit when commanded to do so by others as well. Of course, conditions may range along various dimensions. A dog may reliably sit on command, but only when the ground is dry; it may refuse point-blank to sit in puddles. Or it may reliably sit on command, but only when the command is delivered vocally (hand gestures won't work). Hence, a dog may have a fairly robust disposition to sit on command in some respects (sits in all sorts of weather), but it may be significantly less robust in others (responds only to its master's voice).

Needless to say, robustness in certain respects could matter to us far more than robustness in others, either generally speaking (because certain types of conditions rarely occur) or relative to certain purposes. For instance, we might call a dog "well-trained" so long as it is robustly responsive to its master's commands; being responsive to other people doesn't matter so much because they are not responsible for managing the dog's activities. A similar latitude in attributing dispositional properties can occur with regard to their strength. For instance, under a given range of circumstances, a dog may not be reliably aggressive (disposition to aggression is fairly weak), but the probability of it showing some aggression is not zero; and for some precautionary purposes that may be enough for us to call the dog "aggressive".

Such variability in the nature and attribution of dispositional properties is well noted in the literature. Does this make the attribution of dispositional properties so arbitrary as to undermine their value? Not at all – and for two main reasons. The first is pragmatic. As suggested by the discussion so far, our interest in attributing dispositional properties is deeply connected to how we interact with the things in question. We want to know if the dog is likely to attack, or the vase to break, in the kind of quotidian circumstances that govern our interactions; and we adjust our own behaviour accordingly. Or we may want to know the range of circumstances under which something can be relied upon to 'operate normally'; hence, for instance, we specify the 'load capacity' of an elevator or bridge. More generally, we want to know if things will behave in predictable ways under various types of conditions that have practical significance for us, governing how we interact with them. The attribution of dispositional properties is a short-hand way of marking such practically indispensable information.

The second reason is causal-explanatory. While the attribution of dispositional properties is practically invaluable from the perspective of managing our interactions with the things in question, deeper theoretical concerns direct us to exploring their underlying nature. Assuming a

naturalistic perspective, we want to know how various things are physically constituted so as to support their macro-level dispositional proclivities. The expectation is that such dispositional proclivities are grounded in actual structural features of the things in question, features that are causally responsible for the modally elaborated patterns of behaviour associated with the relevant dispositions -- for instance, molecular structure in the case of the 'fragile' vase. Hence, for many theorists, attributing dispositional properties is simply a way of marking the presence of underlying features on which such properties presumptively supervene, as well as referencing a suitable causal-explanatory theory in which such features are embedded.[1]

Of course, the attribution of dispositional properties may also indicate where more empirical work needs to be done regarding the nature of these grounding features. This is particularly true with regard to the higher-order dispositions of animate creatures: for instance, the dog's disposition to sit on command. While this dispositional proclivity is presumptively grounded in some standing feature of the dog's psychology, itself realized in some complex neural features of the dog's brain, the details of a satisfying causal-explanatory theory are far from complete. And ditto for even more sophisticated higher-order dispositional properties – for instance, our human capacity (or disposition) to respond to various kinds of reasons (e.g., moral or prudential reasons) in the choices we make, whether in theoretical or practical domains.

3. Kinds of dispositional properties

3.1 object-centered dispositions vs. agent-centered abilities

The promise of this framework is compelling for many committed naturalists in the philosophy of mind. In their view, it suggests that there is nothing more to 'being intelligent' in the human sense than possessing a rich array of sophisticated dispositional properties that are somehow realized in the complex architecture of the human brain (call this the ***'basic dispositionalist assumption'***). Furthermore, it is often assumed, at least by many, that such

[1] There are complex issues here that we simply put aside. For instance (and this is by no means an exhaustive list): (1) whether the grounding features of macro-level dispositional properties are themselves dispositions – and perhaps, ultimately, that it's disposition 'all the way down' (for discussion and references, see: Choi and Fara 2021); and (2) whether causal-explanatory theories of macro-level behaviour make essential reference to macro-level dispositional properties -- i.e., these properties are *explanatorily* irreducible in the context of such theories even though they are naturalistically vindicated in virtue of being ontologically grounded in more fundamental features of the things in question (a view that we favour; see, for example: Jackson and Pettit 1990b, 1990a).

sophisticated dispositional properties are really no different in kind from the more basic dispositional properties of other physical objects (call this the ***'deflationary dispositionalist assumption'***). As Kadri Vihvelin representatively summarizes:

"We have the ability to choose on the basis of reasons by having a bundle of capacities which differ in complexity but not in kind from the capacities of things like thermostats, cars, and computers. These capacities are either dispositions, or bundles of dispositions, differing in complexity but not in kind from dispositions like fragility and solubility" (Vihvelin, 2004, p. 429)

Gilbert Ryle (1949) was an early advocate of the basic dispositionalist approach to human intelligence. Like Vihvelin, he argued that the sophisticated qualities of human mentality could be understood in straightforwardly dispositional terms – specifically, as consisting in a multifaceted range of capacities responsible for the distinctively complex patterns of 'overt and covert' behaviour human beings manifest in their day to day lives.[2] But unlike Vihvelin, Ryle staunchly resisted the deflationary dispositionalist assumption, maintaining that an adequate view of human intelligence depended on recognizing the distinctive nature of the dispositions that constitute it. Thus, a key part of his project was to chart the manifest differences amongst kinds of dispositional properties; most obviously, between the dispositional properties of inanimate things (hereafter, '***object-centered dispositions***') and the acquired (learned) dispositional properties of animate creatures (hereafter, '***agent-centered abilities***').

We agree with Ryle that understanding such differences sheds light on the naturalistic underpinnings of human intelligence; but more importantly for our purposes here, we think it also sheds light on the kind of systems (whether artificial or not) that could instantiate the properties in question. We turn now to a consideration of these critical differences, originally highlighted by Ryle, but as explored and elaborated in McGeer (2018).

Begin by considering some prototypical examples of 'object-centered dispositions' versus (acquired) 'agent-centered abilities'. Examples of the former include: fragility, solubility, conductivity. Examples of the latter include: sitting-on-command, tying shoelaces,

[2] Importantly, Ryle's use of the term 'behaviour' is very broad, covering internal mental processes ('covert' behaviour), as well as external bodily activities ('overt' behaviour). Some of his examples of covert behaviour: doing arithmetical sums 'in one's head' or imagining (with trepidation) falling through an ice-covered lake). This feature of Ryle's view is often overlooked. Indeed, he is frequently caricatured as a simple-minded analytic 'behaviourist', who believed that the attribution of mental states is nothing more than the attribution of dispositions to purely external (or 'overt') forms of behaviour. This interpretation is easily refuted by reference to numerous passages in *The Concept of Mind*.

multiplying numbers, playing chess, mountaineering, moral reasoning. Commonsense recognizes an obvious difference between these two classes of dispositional properties. Agent-centered abilities characterize what agents can do in suitable conditions; they are agency-involving 'active powers', essentially linked to the beliefs, desires and intentions of those who possess them. By contrast, object-centered dispositions are 'passive' powers; they simply characterize what various things will undergo in suitable conditions (Moore, 1911; Reid, 1788). Important as this distinction is, it is not one that cuts any ice with naturalistically inclined philosophers of mind. For the dispositional characterization of the psychology of agents that possess the relevant agent-centered abilities is such as to include their proclivity for forming suitable epistemic and motivational states under the conditions in which they manifest their abilities; forming such states is part and parcel of the requisite abilities, not something in addition to them. In short, this commonsense distinction can be accommodated without suggesting that there is any deep difference of kind between these types of dispositional properties; certainly, there is no difference of kind that requires the positing of special agent causal powers to initiate the ability in question ('acts of will', for instance).

Nevertheless, there is another sense in which agent-centered abilities are 'agency-involving' – and this does mark a critical difference in kind between abilities and mere object-centered dispositions. Abilities are actively acquired; they take a distinctive kind of agential work – i.e., ***practice*** – to develop, where practice involves agents intentionally manifesting some approximation of the ability in question and then reshaping how they behave in light of feedback (positive or negative) they receive from the environment. Hence, abilities may be conceptualized as dispositional properties that are constituted by certain intrinsic features of the things that possess them. In this, they are akin to object-centered dispositions – and thus, perhaps, seem merely different in degree of complexity rather than kind. But the crucial difference in kind remains: that agents must actively work to shape the intrinsic features that constitute their abilities (presumptively, these are complex cortical networks); whereas objects do not intentionally work at developing the intrinsic features that constitute their fragility, conductivity, solubility, and so on. They get this for free, simply by virtue of possessing relatively stable intrinsic properties.[3]

How significant is this difference in kind? We said above that the attribution of

[3] This is not to say that the intrinsic properties of objects do not change over time, sometimes by virtue of agents' deliberately acting upon them (e.g., as when glass is 'tempered' via thermal or chemical treatment). Naturally, as the intrinsic properties of such objects change, so will the macro-level dispositional profile such properties realize.

dispositional properties seemingly has nothing to do with how frequently, if at all, the characteristic pattern of behaviour associated with those properties is manifested. They are, as we noted, modal properties – properties that things may possess regardless of the conditions they encounter or undergo in the actual world. This observation makes perfect sense with regard to object-centered dispositions. But it is deeply misleading with regard to agent-centered abilities, miscuing us as to something fundamentally distinctive about their underlying nature. After all, it is only thanks to a continuous and systematic regimen of feedback-driven behavioural approximations of the target ability that it ever comes to be in the first place; and that implies that the relevant manifestations conditions must frequently and consistently occur in the actual world for these properties to exist at all. Thus, it seems, not all modal properties are created equal.

But, again, how significant is this fact? Though it may be empirically salient that agent centered abilities take practice to acquire, qua dispositional properties, they could in principle come to exist in some other way: e.g., by way of clever neural tinkering (or, in the artificial case, programming). After all, if we could create a physical doppelganger of any creature possessed of various agent-centered abilities, then, by the tenets of naturalism, the doppelganger would possess those self-same abilities.[6] But, again, this observation seriously misleads us with regard to the underlying nature of the abilities in question, at least so far as these are instantiated in biological creatures. For the simple fact is that abilities get rusty with disuse, even those that consist in fairly well-entrenched cognitive/behavioural routines. A dog may reliably sit on command; but if its master becomes lax or disappears altogether, and the dog is never again subjected to anything like a consistent reinforcement schedule, the ability to sit on command fades away. Ditto for our much prized and more sophisticated human abilities: speaking a (second) language, riding a bicycle, playing a musical instrument. All of these are subject to practice-deprived decay – a fact that is often obscured to us by the fact that we continually practice many of our most prized abilities on a regular basis (speaking a language, reasoning through problems, riding a bicycle, driving a car). In short, the kind of intrinsic features that realize these sophisticated abilities are critically practice (or manifestation)-dependent, where practice involves an on-going process of deliberately manifesting the relevant behaviour under suitable conditions and receiving appropriate feedback in turn. This makes the modal properties that constitute (biologically based) abilities essentially fragile or ‘labile’. And why is this? Presumptively, because the intrinsic features that realize such properties are dynamically maintained cortical networks: they are networks that require regular activation to maintain integrity.

3.2 Kinds of agent-centered abilities – habits vs. skills (a.k.a. 'Intelligent' capacities)

So far we have charted a critical difference between kinds of dispositional properties – viz. whether they are essentially dynamic, manifestation-dependent modal properties: 'object centered dispositions' are clearly not, whereas agent-centered abilities clearly are (at least as realized in natural biological creatures). But, interestingly, Ryle himself was more concerned with highlighting another distinction – this time, more likely one of degree than of kind – within the class of agent-centered abilities: a distinction that is nonetheless substantial between what he called ***'intelligent capacities'*** or '*skills'* on the one hand, and mere ***'habits***' on the other. Examples of the former include: target-shooting, mountaineering, driving a car, constructing arguments, speaking a language, and moral reasoning. Examples of the latter include, sitting-on command, tying shoelaces, multiplying numbers, reciting the alphabet.

One obvious difference between habits and skills is that habits manifest in a relatively simple, stereotyped pattern of task-appropriate behaviour under a relatively constrained set of behaviour-eliciting conditions. By contrast, skills manifest in a much richer pattern of task appropriate behaviour, under a relatively expanded and unconstrained set of behaviour eliciting conditions. As Ryle says, skills are dispositional properties "… the exercises of which are indefinitely heterogenous" (1949, p. 44). Hence, even if habits and skills both require a kind of agential work – viz. practice – to develop and maintain, the dispositional outcome seems to be rather different in kind. What could explain this difference? There are in fact three dimensions of difference to which Ryle draws attention, and we discuss each of these in turn.

The first concerns the ***kind of work*** it takes to develop and – in particular— 'maintain' the ability in question. Habits require practice, that goes without saying. They take time to develop and suitably hone. But once established, they consist in entrenched cognitive and behavioural routines that agents are able to reproduce without much conscious attention or effortful monitoring. As Ryle observes, "when we describe someone as doing something by pure or blind habit, we mean that he does it automatically and without having a mind to what he is doing" (1949, p. 42). By contrast, skills are often exercised in a conscious, attentive way. Skilled agents must often 'think what they are doing' as they engage in the relevant activity, shaping their behaviour with 'care, vigilance and criticism', mindful of problematic circumstances they encounter on the ground. And even when they are operating in a relatively unreflective mode, they must be ready to tune into such problematic circumstances were they

to appear. In short, they must be attentive (whether in virtual or active mode) to how things might go wrong in exercising their skill – hence, with on-going receptivity to potentially corrective feedback received from the environment (ibid., p. 42).

But why should all this be necessary? Surely once a degree of proficiency is reached, the activities of skilled agents must likewise depend on their having entrenched an appropriate range of cognitive and behavioural routines. So why not 'go on auto-pilot'? The answer is straight-forward. The task-appropriate behaviour that constitutes the exercise of a skill is generally more demanding than the task-appropriate behaviour constituting a habit. It is more demanding because the essence of skilled behaviour is that it is flexible and adaptive; it can be 'reproduced' under new and challenging circumstances – so, in a sense, is not reproduced at all, but rather modified anew as circumstances require. And this means skilled agents cannot simply rely on entrenched cognitive and behavioural routines; they must be ready to adapt such routines to cope with unexpected difficulties and/or novel situations, whether 'on the fly' or over time, in response to suitable feedback from the environment. As Ryle says,

"it is of the essence of merely habitual practices that one performance is the replica of its predecessors. It is of the essence of intelligent practices that one performance is modified by its predecessors. The agent is still learning" (Ibid., p. 42).

The second dimension of difference between habits and skills concerns the ***kind of feedback*** agents require to develop and maintain these different kinds of abilities. Ryle speaks here of 'drilling' vs. 'training'. These terms are not perhaps ideal, but what Ryle has in mind is that certain kinds of feedback aim to inhibit creative responsiveness, whereas other kinds of feedback aim to promote such responsiveness. He characterizes the difference between them as follows:

"Drill (or conditioning) consists in the imposition of repetitions…. Training, on the other hand, though it embodies plenty of sheer drill, does not consist of drill. It involves the stimulation by criticism and example of the pupil's own judgement. He learns how to do things thinking what he is doing, so that every operation itself is a new lesson to him on how to perform better… Drill dispenses with intelligence, training develops it" (Ibid., pp. 42-42).

It is difficult to operationalize how different kinds of feedback might accomplish such different cognitive/behavioural outcomes. But one thing is clear: in developing the creative, intelligent responsiveness that characterizes a skilled agent's profile of task-appropriate adaptive behaviour, the agent must call upon internal resources for judicial review and assessment of feedback from the environment, not to mention a readiness to 'try new tricks',

themselves judiciously weighed in term of probable success and downside risks. In this sense an agent becomes 'responsible for' their performance in a distinctive kind of way – distinctive because we are now in the domain of 'judgment-sensitive' attitudes and behaviour: attitudes and behaviour that are reason-governed, and not simply reason-conforming; hence, open to the agent's reflective review and control (for representative discussion, see: Pettit and Smith 1996; Scanlon 1998; Smith 2008).

The third dimension of difference between habits and skills concerns their ***underlying nature***. As we have noted thus far, they both take agential work to develop; and they both take agential work to maintain – i.e., habits and skills both get rusty with disuse. By our lights, they are thus both dynamically maintained, manifestation-dependent modal properties. Yet there is nonetheless an important difference between them (perhaps a difference merely of substantial degree). To bring this out, consider that feature of modal properties we earlier called 'robustness'. To recapitulate: robustness concerns the range of conditions under which the thing in question would manifest the disposition-relevant behaviour: the wider the range of such conditions, the more robust the dispositional property. Habits and skills clearly differ in this regard: not only do habits involve constrained, stereotyped forms of task-appropriate behaviour, they are exercised under a relatively restricted range of conditions. But not so with skills: the range of conditions under which they are exercised is far more extensive and variable, as is the form of (adaptive) behaviour appropriately elicited under those wide ranging conditions. To repeat Ryle's observation, they are dispositional properties "… the exercises of which are indefinitely heterogeneous".

But now we face something of a puzzle. As dispositional properties, both habits and skills are presumptively realized in dynamically maintained cortical networks. These are the intrinsic features of agents on which such properties supervene. But it defies credibility to imagine that skilled agents could be, here and now, intrinsically so structured that they are, as it were, primed to manifest the 'indefinitely heterogenous' forms of task-appropriate behaviour under the equally heteroneous range of conditions in which a skill is generally exercised. In short, from a naturalistic design perspective, skills seem to be nothing short of miraculous.

This puzzle is easily solved, however, if we remind ourselves that skills are the kind of dispositional properties that are continuously undergoing development; hence, the 'robustness' that characterizes them is, in effect, spread out over time. Agents are able to produce relevant instances of task-appropriate behaviour under such wide-ranging conditions because they are continuously adjusting the dynamic cortical networks that presumptively underlie the successful performance of their skilled behaviour. In sum, to possess a skill agents

must be psychologically so structured that, at any given time, they are primed to engage with the world, creatively, flexibly and judiciously, in ways that drive their own development.

So finally the differences amongst dispositional properties comes down to this. At the abstract level, they are certainly all modal properties. But modal properties differ in significant ways; and this is relevant to the question of the kind of systems in which they might be instantiated. In the first place, there are modal properties that exist only by virtue of being dynamically instantiated through deliberate agential activity – these are 'manifestation-dependent' modal properties (habits and skills). But in the second place, there are two types of such properties - varying substantially in degree, if not in kind: those that are *dynamically sustained* through practice (habits) and those that are *dynamically-revised* through practice (skills). Or to put this difference another way: habits are modal properties that maintain their stability by way of an agent's (feedback-sustained) repetitive practice; skills are modal properties that maintain their open-textured integrity by way of an agent's (feedback-soliciting) intelligent practice, where such practice is judiciously guided by the agent's on-going sensitivity to the adaptive requirements of novel or changing circumstances.

4. Artificial systems and their dispositional properties

In this section, we look into the nature of dispositional properties in the context of artificial systems. In particular, we ask to what extent could the dispositional properties explored in Section 3 be instantiated in artificial systems. We focus on two paradigmatic kinds of artificial systems: rule-based classical AI and reinforcement learning AI. We argue that these artificial systems are not equal in the kinds of dispositional properties they instantiate. In particular, rule-based classical artificial systems instantiate merely object-centered dispositions, whereas reinforcement learning systems allow for the instantiation of something like agent-based abilities.

4.1. Classical Rule-based artificial systems

Classical rule-based AI — or "Good old-fashioned artificial intelligence" (GOFAI) — is a term of art dubbed by John Haughland (1985). GOFAI was the dominant paradigm of AI research from the 1950s to the 1990s. GOFAI systems take intelligence to be encompassed in the logical manipulation of symbols. A symbol is an item of a formal language which can be merely regarded as a purely syntactic structure or interpreted as representing explicit and

discrete objects or events in the world.

A GOFAI system is an interconnected network of such discrete symbols with logical inference rules applied to them. This system employs specific logical rules (programmed into the system) on formal symbolic representations, constructs and transforms the symbolic data structures according to different rules for searching and planning the symbols. A successful GOFAI system follows its programmed logical rules and draws logical inferences when performing particular tasks.

An early example of a GOFAI artificial system, ELIZA, was developed by Joseph Weizenbaum (1966). ELIZA was a natural language processing program initially designed to simulate a human psychotherapist. It 'engaged in conversation' with a human user by way of manipulating the symbolic representation of human responses according to a set of pre-programmed logical rules. However, ELIZA's capacity to mimic a human conversation was highly constrained. Its responses were rigidly determined by the specific input it received and the local rules for manipulating or transforming that input to an output. Consequently, ELIZA lacked the flexibility to adapt its responses beyond the strict pre-determined rules and inputs, a significant limitation that restricts the scope of its conversational capabilities.

By definition, GOFAI systems produce their seemingly intelligent behavior by drawing classical logical inferences through the manipulation of symbols. Their capacities are dependent on the inferential nature of the rules of classical logic. As long as these rules are instantiated in the system, they can follow them and draw inferences from them. But now we ask, what kinds of dispositional properties do GOFAI systems instantiate according to the taxonomy defended in Section 3?

At first glance, the dispositions of GOFAI systems may seem rather unlike the object-centered dispositions of physical objects earlier described, such as fragility or conductivity. We tend to conceive of GOFAI systems as computer programs or software that can be implemented on a diverse range of materials, ranging from silicon to bio-inspired materials. Nevertheless, even though these systems primarily function in the form of logic-driven code rather than physical, embodied entities, they possess properties that mirror those of object-centered dispositions. In much the same way as a glass object's fragility is a stable feature, the inferential properties of GOFAI systems are enduring attributes, as they are integral components of the entities that contain them. These logical rules are deeply embedded in the fabric of the GOFAI systems, and under characteristic conditions, do not adjust to pressure

of feedback from the environment.

For example, consider a GOFAI system designed to play chess. It is pre-programmed with specific logical rules, ensuring the system will exploit certain strategies geared to specific board configurations. This is the system's object-centered disposition. Just as the molecular structure of a glass object makes it fragile, the pre-set rules and strategies give the chess-playing system its ability to play chess. This object-centered feature is an intrinsic, stable feature of the system under characteristic conditions.

While GOFAI systems instantiate object-centered dispositions, they do not have the right sort of structure to instantiate agent-centered dispositions as we have described them. These systems are not involved in any form of active learning; nor do they maintain their dispositional properties by way of active engagement with the environment. While GOFAI systems do require programming in order to 'acquire' their dispositional properties, the programming itself is not self-initiated. If the GOFAI chess system is repeatedly defeated by a particular strategy from an opponent, it will not adapt or develop new counter-strategies. Hence, once in place, the programmed structure and dynamics of a GOFAI system are stable intrinsic features of that system, analogous to the molecule structure of a fragile glass. In sum, while GOFAI systems may display an impressive range of complex behaviour thanks to the dispositional properties they instantiate, they are nonetheless not 'intelligent' in our proprietary use of this term; they are not actively engaged in developing and/or maintaining their own dispositional proclivities.

4.2. Reinforcement-Learning artificial systems

The second kind of artificial systems we discuss is the reinforcement learning (RL) system. RL is a broad class of artificial systems that focuses on training agents to maximize their rewards within a given environment (Sutton and Barto, 2018). The primary characteristics of this paradigm of artificial systems are the agent and the environment. The overall goal of an RL agent is to maximize its total reward (or minimize its total punishment) while interacting with the environment. At each step of interaction with the environment, the agent makes an observation of the state of the world, which is often partial due to the inherent limitations in its ability to fully perceive the environment, and accordingly decides on an action to take. The agent learns its path towards the maximization of its total reward by using feedback from the environment (rewards or penalties) contingent upon every action it performs.

Rewards and penalties are signals from the environment that tell the agent how good or bad the current state of the world is. An RL agent is not directly endowed with any logical rules from a human programmer that guide it to the maximization of the total reward. Instead, an RL agent learns in interaction with an uncertain environment by starting from random trials and finishing with sophisticated tactics for resolving the problem it is set to achieve by leveraging the power of trial and error.

The RL approach has been successfully applied to a wide range of tasks, from playing games to controlling robots and self-driving cars. RL systems possess various degrees of capabilities. For example, some of them have learnt how to play complex strategy games such as the game of Go (Silver et al., 2016). Go has been traditionally considered one of the most challenging games for computer programs to master due to its large search space and the need for strategic thinking. In 2016, AlphaGo – an RL system developed by Google DeepMind – defeated Lee Sedol, one of the world's top Go players, in a five-game match.

What kinds of dispositional properties do RL systems instantiate? Unlike GOFAI, reinforcement learning AI systems are designed in such a way that they need to receive feedback through a dynamic connection to the environment in order to acquire their task-specific abilities. For example, AlphaGo uses a combination of deep learning and Monte Carlo tree search to make its decisions. The deep learning component, called the "value network," estimates the probability that a given move will lead to a win. The Monte Carlo tree search component, called the "policy network," uses this information to search the tree of possible moves and select the best one. One of the key features of AlphaGo is its use of supervised learning to train the value network. This network was trained on a dataset of over 30 million expert Go moves, which allowed it to learn the patterns and strategies of top players.

To emphasize the diversity of RL systems, let us look into their incorporation in a complicated technology: their use in self-driving cars (e.g., Waymo). By definition, self-driving cars should be capable of sensing their environment and navigating it without constant acquisition of human input. For their navigation, they use a variety of technologies, such as radar, lasers, and cameras, to sense their surroundings and make appropriate decisions. One of the challenges in developing self-driving cars is the need to train the AI agents that control the vehicle to make decisions based on their environment. This is where reinforcement learning comes in. In the case of a self-driving car, the AI agent may be trained to avoid obstacles, such as other vehicles or pedestrians. If the AI agent successfully avoids an obstacle, it will receive a reward. If it fails to avoid the obstacle, it will receive a punishment. Over time, the AI agent *learns* to make decisions that maximize its rewards and minimize its

punishments. This means that the AI agent can adapt to changing environments and make more accurate decisions.

Despite some architectural differences and nuances, an RL agent needs to work towards acquiring its dispositional properties; hence, it manifests what we call 'agent-centered' abilities. This is a point that we will elaborate on in the next section.

5. The blended agent-centered abilities of reinforcement learning systems

So far, we have argued that, unlike GOFAI, reinforcement learning artificial systems acquire agent-centered abilities. In this section, we argue that RL systems can be instantiated in various ways, embodying different degrees and types of learning and adaptability to their environments. In Section 3, we suggested that, in the human case, there is an important difference between habits (routinized, relatively inflexible patterns of behaviour) and 'intelligent capacities' or skills (patterns of behaviour that require adaptive sensitivity to complex and potentially changing conditions on the ground). Although this difference is undoubtedly one of degree more than of kind, it is an important one to mark; and in this section we consider how this distinction might apply in the case of RL systems.

One possibility is that the distinction maps cleanly into the domain of reinforcement learning systems. Specifically, given the simplicity of tasks less advanced RL systems are trained to perform, the abilities they instantiate share much in common with habits. Examples include navigating simple mazes (Mnih et al., 2015) or optimizing basic control tasks like balancing a pole (Schulman et al., 2015). More advanced RL systems, on the other hand, operate in more complex environments; hence, the abilities they require share more in common with skills in so far as they must produce a more flexible pattern of behaviour in order to perform adequately given the task demands. Examples include playing complex games like Go (Silver et al., 2016) or chess (Silver et al., 2017), generating human-like text (Radford et al., 2019), or making recommendations based on user preferences (Zheng et al., 2018).

We agree that something like this distinction applies in the case of RL systems. But the picture is complicated by an underlying fact that we have yet to address – viz., the significant ways in which the agent-centered abilities of these artificial systems differ more systematically from human abilities, producing an overall difference in quality that crosscuts the habits-skills distinction as applied in the human case. These differences span four dimensions: acquisition requirements, maintenance requirements, reliability in manifestation, and finally adaptability to new circumstances (earlier identified as the hallmark of human skills). We discuss each of these

dimensions, before turning in the next and final section to consider what significance this may have for the larger topic of artificial responsible agency.

First dimension of difference, acquisition requirements. Less advanced RL systems learn exclusively through training data sets, in stark contrast to the more organic, in situ practice that humans engage in. These systems are typically fed large amounts of high-dimensional data, which they then use to discern patterns and make predictions. A major concern, however, is the problem of feature selection within these data sets. It can be challenging to ascertain exactly what features an RL system is picking up on, leading to proficiency within the training set that may not necessarily translate well to real-world prediction scenarios.

More advanced RL systems, on the other hand, exploit more dynamic and adaptable learning methods that mirror some of the complexity found in human learning. These systems can leverage techniques such as transfer learning (Torrey and Jude, 2010) or meta-learning (Vanschoren, 2019) to generalize from one task to another, or to quickly adapt to new tasks, respectively. This reduces the dependency on specific training sets and improves the system's ability to handle novel, real-world scenarios. However, these advanced systems still have their limitations and cannot fully replicate the fluidity and adaptability of human learning.

For humans, learning occurs both implicitly and explicitly within complex environments that include simulated scenarios, training data, and real-world situations. The diversity and complexity of these learning environments often surpass what current RL systems can handle. Furthermore, human pedagogic practices are crafted to deal with expectable errors in apprentice learning, producing structured learning environments that are specifically geared to human (embodied) cognitive propensities (Sterelny, 2012). This is a level of foresight not yet fully realized even in current advanced RL systems.

However, both less advanced and more advanced RL systems do share a distinct advantage over humans: the ability to transfer learned skills quickly and all-in-one-package to other systems, e.g., by duplicating the RL systems. This eliminates the need for each RL system to undergo the same training regimen. However, to continue advancing and adapting, these RL systems must maintain active learning capabilities.

Second dimension of difference, maintenance requirements. Maintaining proficiency in specific abilities differs notably between humans and RL systems, sometimes seemingly giving the upper hand to RL systems. Human abilities necessitate ongoing practice to avoid degradation, as our biological systems are inherently adaptive and require constant reinforcement. Consider musicians, who must regularly practice their instrument to maintain their performance level, or

athletes, who need consistent training to uphold their physical prowess and technique.

In contrast, RL systems, which function through algorithmic frameworks such as deep neural networks, do not require the same level of continuous maintenance once they have mastered the necessary abilities according to their designers. Learned information is primarily hard-wired within their algorithmic structure, making them less vulnerable to proficiency loss over time. Hence, the performance of an RL agent remains relatively stable even in the absence of regular practice.

Interestingly, the very aspect that requires humans to consistently practice to maintain their abilities might also confer a significant advantage. Our cortical networks, on which these abilities supervene, are continually fine-tuned to accommodate both minor and major changes in our circumstances. This adaptability may explain the resilience of human abilities across a wider range of real-world conditions than is typically found in RL systems – a point we will come back to below.

Third dimension of difference, reliability in manifestation. As noted in Section 1, dispositional properties, including human abilities, differ in how reliably they manifest under characteristic conditions, ranging all the way (at least in principle), from 'sure-fire' dispositional properties (they always manifest under characteristic conditions) to those that are only weakly reliably (episodically manifesting under characteristic conditions).[4] We called this feature the 'strength' of a dispositional property. But the reliability of a dispositional property can also be affected by its 'robustness' – i.e., the range of characteristic conditions (narrower or broader) under which it can be expected to manifest. On this measure, the broader the range of such conditions, the more reliable will be the dispositional property in its manifestation.

Turning now to human abilities, their reliability is highly dependent on training and practice. In general, the more ingrained the habit, or the more developed the skill, the more reliable their manifestation under suitable conditions. Nevertheless, whatever their stage of development, all human abilities are subject to a number of contingently occurring internal factors that can derail their characteristic manifestation – e.g., fatigue, distraction, cognitive overload. A tired driver might not react as swiftly as when they are well-rested, a distracted pianist might hit wrong notes during a concert, and a stressed or anxious tennis pro may repeatedly botch their serve in an important tennis tournament.

By contrast, RL systems might seem potentially more reliable in manifesting their abilities

[4] By 'characteristic conditions', we simply mean those conditions under which the dispositional property would typically manifest.

under suitable conditions in so far as they avoid the weaknesses that plague human agents. But they suffer from their own constraints in this respect, depending on the type of RL system in question.

Less advanced RL systems may exhibit high reliability in environments that closely mimic their training conditions – this is akin to having a very strong, perhaps even sure-fire dispositional property. However, such systems may fall apart in conditions that deviate only mildly from these training conditions; moreover, such failures are not easily predicted, since the conditions under which they may falter are generally not those that would derail the performance of a human agent. In short, despite the s*trength* with which they are likely to manifest, the abilities of less advanced RL systems are unpredictably narrow in *robustness* thanks to constraints in their learning algorithms that prohibit generalization to new or altered conditions. And this may create serious reliability issues when it comes to anticipating the real-world environmental conditions under which their abilities will be manifested.

More advanced RL systems, meanwhile, are designed to overcome this limitation by exhibiting a higher degree of adaptability, leveraging sophisticated algorithms that allow them to learn from novel situations and adjust their behaviour accordingly. However, these systems are not without their own constraints. For instance, they may require significantly more computational resources to operate effectively, analogous to a human needing more mental or physical energy to perform a task. And when those resources are limited, they may perform sub-optimally – becoming unreliable in the manner or a fatigued, distracted or stressed human being. Additionally, such systems may struggle when faced with situations that drastically deviate from their training environments, just as human beings may struggle when faced with completely unfamiliar scenarios. Hence, despite such algorithmic advances, these systems still face challenges in consistently manifesting their skills in a reliable manner, especially in dynamic or unexpected environments.

Fourth dimension of difference, adaptability. We come at last to what we have called the hallmark of possessing a human-like skill or 'intelligent capacity' – the agent's ability to adapt their behaviour to meet new or challenging conditions by way of leveraging their reservoir of previous experiences and their advanced cognitive faculties. For instance, consider a person who is relatively skilled in driving a car. This individual can typically adapt their behaviour, with minimal practice and corrective feedback, to a spectrum of vehicles, variable road circumstances, and even disparate traffic regulations (such as driving on the right or left). This inherent versatility paves the way for swift and efficient adaptation, distinguishing humans from numerous other species.

By contrast and as we have already noted, less advanced RL agents frequently struggle with the task of generalizing their acquired abilities to different environments or circumstances, even when those abilities involve a relatively complex suite of responses often associated with human skills. Hence, they can achieve remarkable proficiency within a specific scope, but their expertise is commonly confined to their training goals. In this, their abilities are more like habits than skills.

However, more advanced RL systems are starting to bridge this gap. These sophisticated models, often equipped with transfer learning capabilities, show promise in their ability to apply knowledge and skills from one domain to another. Of course, this process often requires additional training or fine-tuning in the new context; and they are still far from matching the natural flexibility and adaptability of human cognition.

In sum, while less advanced RL agents have clear limitations in this dimension, more advanced systems are making strides towards greater adaptability. Nevertheless, the natural adaptability inherent in humans remains a challenging benchmark for artificial intelligence systems to meet. The quest for truly versatile and adaptable artificial intelligence continues to be a fundamental challenge in the field.

6. From artificial dispositions to 'responsible agents'?

With the advent of more advanced AI systems, potentially operating with greater and greater autonomy in high-stakes (potentially harm-causing) settings, theorists have become increasingly concerned with the question of whether such systems might be or become genuinely responsible for what they do – i.e., responsible in the manner of ordinary human agents. This is a complex question, largely driven by a concern with so-called 'responsibility gaps' (Matthias 2004; Purves et al, 2015; Himmelreich 2019; Gunkel, 2020; Santoni and Giulio Mecacci, 2021; Tigard 2021).

A responsibility gap is purported to occur when: (a) a system is operating as a self-standing intentional agent, making choices for which a typical human agent would be held responsible, but is seemingly not a fitting target of our responsibility practices; and (b) no human agent is suitably connected to the system's choices to be responsible for them either.[5] 'Responsibility gaps' are bad news, on most theorists' views, because they undermine the very terms in which many of our

[5] At best, human agents would be responsible for designing and/or delegating control to the system in question, but that is not the same as having direct agential responsibility for the choices made.

social arrangements are structured – viz. the fact that we are able to hold one another to account (whether legally or morally) for harm-causing acts and omissions. Of course, existing social arrangements do incorporate agents that are not responsible for what they do (e.g. young children, the deeply cognitively disabled). But these arrangements ensure that such agents are excluded from so-called 'positions of responsibility' – paradigmatically, high-stakes (potentially harm-causing) settings in which they are allowed to operate autonomously.

How do theorists respond to the threat of responsibility gap caused by AI systems? For those who are pessimistic about the prospects of artificial responsible agency, there are essentially two choices canvassed in the literature: (a) simply resist the allure of deploying autonomous AI systems in these settings (Sharkey 2020; Johnson and Miller 2008); or (b) deploy such systems, but acknowledge that significant changes will be required in our accountability practices to cope with the resultant responsibility gaps when things go wrong (for instance, extend the depth and reach of strict liability law, penalizing those who develop and deploy AI systems for actions and omissions that are not directly sourced in their own agency) (Matthias 2004).

Others are more optimistic about the prospect of designing systems that are functionally equivalent to responsible human agents (e.g., List 2021): this means, not only in their operation (the range of considerations to which they are sensitive), but also in being an appropriate or fitting target of our accountability practices. If so, then responsibility gaps could be avoided more directly. But serious questions remain as to what such functional equivalence would really entail.

Our own contribution to this debate has been comparatively modest and indirect. Instead of confronting the problem of artificial responsible agency head-on (where this generally implies having an appropriate sensitivity to moral concerns), we have focused on a more basic phenomenon that we hope may shed some light on what such agency might entail. This more basic phenomenon concerns the instantiation of agent-centered abilities, and specifically those that share much in common with human-like skills.

As we have argued, skills constitute 'intelligent' capacities in so far as they are relatively flexible and adaptive as compared with mere habits. For human agents, this implies that they only operate in a skilled way insofar as they are generally prepared to adjust how they behave to suit the situations they encounter, whether that be *ex ante* (by way of anticipating specific problems) or *ex post* (by way of responding to corrective feedback). But what does this have to do with 'responsible agency'? Ryle suggests, and we follow him in this, that skilled human agents are responsible for what they do in one straightforward sense of that term. They are able to ***take responsibility*** for shaping and regulating their own behaviour in skill-relevant ways. As Ryle says,

> *"To be intelligent is not merely to satisfy criteria, but to apply them; to regulate one's actions and not merely to be well-regulated. A person's performance is described as careful or skilful, if in his operations he is ready to detect and correct lapses, to repeat and improve upon successes, to profit from the examples of others and so forth. He applies criteria in performing critically, that is, in trying to get things right" – CoM, pp. 28-29*

In Ryle's conception, skilled agents are not simply sensitive to features of situations that bear on how well or badly they are performing in skill relevant ways, they have some understanding – indeed, as they become more skilled, a better understanding –of what those features are. In short, they are becoming sensitized to those features as *reasons* for governing their behaviour one way or the other. And this in turn explains the impressively open-ended adaptability of skilled behaviour in the human sense. For in becoming reasons-responsive, such agents are able to tune into features of novel circumstances as potentially reason-giving; they are able to profit from the reason-giving examples and feedback of others; and they are even able to explain what they are doing, whether successful or not, in reason-giving ways, thereby becoming explicitly answerable for what they do.

To our mind, it is an open question whether RL-systems that approach the natural adaptivity of human skilled behaviour have what it takes to be reasons-responsive – and thus responsible for their behaviour – in this sense. On one hand, with simpler RL systems it's certainly possible that they can demonstrate sensitivity to their environment and actions without truly understanding the reasons behind their responses. For instance, an AI trained to play a game might improve its performance by learning from its mistakes and adapting its strategy. But does it *understand* the reason why one strategy works better than another, or is it merely optimizing its actions based on the rewards it receives? On the other hand, with more complex systems, the answer may be less clear. At what point does their increasingly sophisticated sensitivity shade into genuine reasons-responsiveness? Is this a difference of degree or a difference of kind? We think it is plausible that we are talking here of a difference of degree; or, rather, that increasing sensitivity of the requisite complexity will necessitate AI becoming truly reasons-responsive, capable of conceptualizing (i.e., explicitly representing) as well as processing the reasons for its actions. This would imply a level of understanding and cognition that goes beyond simple reaction to stimuli. If we aim for AI systems to be accountable for their predictions or decisions, they must exhibit this capacity for reasons-responsiveness. This is crucial for our ability to trust them. The field of explainable AI is dedicated to this pursuit, seeking to create AI systems that can provide understandable explanations for their predictions and decisions (Gunning et al., 2019; Miller,

2019; Günther and Kasirzadeh, 2021).

In our estimation, if the skilled RL -systems of the future have what it takes to be reasons-responsive, then *ipso facto* they count as 'responsible agents'. For they are responsible in the most basic sense of that term: they shape and regulate their behaviour according to their understanding of how they *ought* to behave in order to perform successfully; they are able to provide reasons for their behaviour; and they are able to understand corrective feedback (whether from the environment or from others) as sensitizing them various reasons for their past failures (i.e., skill-relevant features of the situation they ought to have taken into account, but didn't).

But this is not yet to say that they count as *morally* responsible agents. Admittedly, when theorists talk about 'responsible agents', this is often what they have in mind – agents that are responsive to specifically ***moral*** reasons, considerations that bear on whether acts and judgements are morally acceptable or unacceptable, morally better or worse. Still, if skilled RL systems have what it takes to be reasons-responsive, we are one step closer to genuine moral agency. For, on many compatibilist views, it is (moral) reasons-responsiveness that lies at the core of such agency.

So this leads to another set of open questions: could the skilled RL-systems of the future become reasons-responsive in this more specialized sense (cf. Purves et al, 2015)? What would it take to sensitize such systems to moral considerations, where these provide added constraints on skilled performance.? Are there grounds for thinking that this cannot be accomplished in precisely the same way that AI systems come to have intelligent capacities in the first place – for instance, by the provision of appropriate corrective feedback in the face of morally inadequate behaviour? Or is there something special about human beings – for instance, perhaps our affective nature – that makes us peculiarly sensitizable to moral considerations that cannot be functionally mimicked or replaced in AI systems?

A final range of open questions follows from this. Suppose future RL-systems are sensitizable in adequate measure to moral considerations, constraining their behaviour in morally acceptable ways. Does this make them fit and proper targets of our accountability practices? Would the operation of such systems in high-stakes (potentially harm causing) settings leave us with an unacceptable 'responsibility gap' when things go wrong? Or is there some meaningful way in which we could hold such systems to account?

In our estimation, such questions cannot be addressed without some consideration of what our accountability practices are for; and, of course, this is a much-debated topic. But there is one tradition of thought that gibes rather well with their playing a critical role in sensitizing AI systems to the relevant moral considerations – viz., providing the kind of corrective feedback that enables

AI systems to improve their performance going forward.[6] Now many will argue this does not go to the heart of our accountability practices: their raison d'etre is not to improve moral performance, but rather to punish offenders or at least to instill in them some pained recognition of the harm they have caused, prerequisite – or so it is thought – for deepening the offenders' moral understanding, as well as assuaging the victim's own pain and suffering (Danaher, 2016).

Again, there is much to address in these concerns; and in so far as they seem compelling, there is little prospect of the AI systems we envision in this paper being suitable targets of our accountability practices. For these seemingly all depend on the affective dimension of our moral lives – in particular, that human beings suffer, or can be made to suffer, as a morally fitting consequence of (culpably) harming others. But, again, it seems worth asking why do we care so much that wrongdoers suffer in consequence of the wrong they do? Could there be some functional rationale to this sentiment – perhaps, for instance, that we see wrongdoers as harbouring a *motivational* deficit in their dealings with others (e.g., a lack of care) that we hope to rectify by making their wrongdoing affectively salient to them? If so, then it seems that, in dealing with AI systems, we could abandon this dimension of our accountability practices without any functional loss. Is that sufficient to assuage the depth of our normative concern with 'responsibility gaps' that may result from the use of (suitably sophisticated) AI systems in high-stakes settings? Undoubtedly not. But it may help bring those concerns into sharper focus.

References

Choi, Sungho, and Michael Fara. 2021. "Dispositions." In *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta.

Danaher, John. 2016. 'Robots, law and the retribution gap', *Ethics and information technology*, 18: 299-309.

Gunkel, David J. 2020. 'Mind the gap: responsible robotics and the problem of responsibility', *Ethics and information technology*, 22: 307-20.

Günther, Mario, and Atoosa Kasirzadeh. "Algorithmic and human decision making: for a double

[6] Many will associate such a view with the utilitarian, deterrence-oriented 'optimists' derided by PF Strawson in his famous essay, "Freedom and Resentment" (Strawson 1974; For representative optimists, see: Nowell-Smith 1948; Schlick 1939; Smart 1961)). But there are less instrumental ways of developing such a view. For instance, it might be argued that 'improving moral performance' requires the development of an agent's moral understanding (i.e., sensitization to moral considerations). Hence, providing 'corrective feedback' would have this as its primary goal and rationale. Cf. Adam Smith: "To bring him back to a more just sense of what is due to other people, to make him sensible of what he owes us, and of the wrong that he has done to us, is frequently the principal end proposed in our revenge, which is always imperfect when it cannot accomplish this." (Smith 1759/1982: III, ii, 1).

standard of transparency." *AI & SOCIETY* (2022): 1-7.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. Z. (2019). XAI—Explainable artificial intelligence. *Science robotics*, *4*(37), 17–19.

Himmelreich, Johannes. 2019. "Responsibility for killer robots", *Ethical Theory and Moral Practice*, 22: 731-47.

Haugeland, J. (1985). *Artificial Intelligence: The Very Idea.* MIT Press.

Jackson, Frank, and Philip Pettit. 1990a. "Causation in the Philosophy of Mind", *Philosophy and Phenomenological Research*, 50: 195-214.

———. 1990b. "Program Explanation: A General Perspective", *Analysis*, 50: 107-17.

Johnson, Deborah G, and Keith W Miller. 2008. "Un-making artificial moral agents", *Ethics and information technology*, 10: 123-33.

LeCun, Y; Bengio, Y; Hinton G. (2015). "Deep Learning." *Nature* 521(7553): 436–444.

Legg, S., & Hutter, M. (2007). "Universal Intelligence: A Definition of Machine Intelligence. *Minds and Machines*, *17*(4), 391-444.

List, Christian. 2021. "Group agency and artificial intelligence", *Philosophy & Technology*, 34: 1213-42.

Matthias, Andreas. 2004. "The responsibility gap: Ascribing responsibility for the actions of learning automata", *Ethics and information technology*, 6: 175-83.

McGeer, V. (2018). XV—Intelligent Capacities. In *Proceedings of the Aristotelian Society* (Vol. 118, No. 3, pp. 347-376). Oxford University Press.

Miller, Tim. "Explanation in artificial intelligence: Insights from the social sciences." *Artificial intelligence* 267 (2019): 1-38.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. "Human-level Control through Deep Reinforcement Learning." *Nature*, 518(2015), 529–533.

Moore, G.E. 1911. *Ethics* (Oxford University Press: Oxford).

Nilsson, N.J. (1991). "Logic and Artificial Intelligence." Artificial Intelligence, 47(1), 31--56.

Nowell-Smith, P. 1948. "Freewill and Moral Responsibility", *Mind*, 57: 45-61.

Pettit, Philip, and Michael Smith. 1996. 'Freedom in Belief and Desire', *Journal of Philosophy*, 93: 429-49.

Purves, Duncan, Ryan Jenkins, and Bradley J Strawser. 2015. 'Autonomous machines, moral judgment, and acting for the right reasons', *Ethical Theory and Moral Practice*, 18: 851-72.

Reid, Thomas. 1788. *Essays on the Active Powers of Man* (Bobbs-Merril: Indianapolis).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2019). Improving language understanding by generative pre-training. OpenAI Blog, 1(8).

Ryle, G. (1949). *The Concept of Mind*. Chicago: University of Chicago Press.

Santoni de Sio, Filippo, and Giulio Mecacci. "Four responsibility gaps with artificial intelligence: Why they matter and how to address them." *Philosophy & Technology*

34 (2021): 1057-1084.

Scanlon, T.M. 1998. *What We Owe To Each Other* (Harvard University Press: Cambridge, Mass).

Schlick, Moritz. 1939. *The Problem of Ethics* (Prentice-Hall: New York).

Schulman, J., Levine, S., Abbeel, P., Jordan, M., & Moritz, P. (2015). Trust region policy optimization. In *International Conference on Machine Learning* (pp. 1889-1897).

Sharkey, Amanda. 2020. “Can we program or train robots to be good?”, *Ethics and information technology*, 22: 283-95.

Silver, David, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, Yutian Chen, Timothy Lillicrap, Fan Hui, Laurent Sifre, George van den Driessche, Thore Graepel, and Demis Hassabis. 2017. 'Mastering the game of Go without human knowledge', *Nature*, 550: 354-59.

Smart, J. J. C. 1961. 'Free-Will, Praise and Blame', *Mind*, 70: 291-306.

Smith, Adam. 1759/1982. *The Theory of the Moral Sentiments* (Liberty Classics: Indianapolis).

Smith, Angela M. 2008. 'Control, responsibility, and moral assessment', *Philosophical Studies*, 138: 367-92.

Smith, B. C. (2019). The Promise of Artificial Intelligence: Reckoning and Judgment. MIT Press.

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

Strawson, P.F. 1974. 'Freedom and Resentment.' in, *Freedom and Resentment and Other Essays* (Methuen: London).

Sutton, R.S., and Barto, A. G. (2018). Reinforcement Learning: An Introduction. MIT Press.

Tigard, Daniel W. 2021a. 'Artificial moral responsibility: How we can and cannot hold machines responsible', *Cambridge Quarterly of Healthcare Ethics*, 30: 435-47.

Torrey, Lisa, and Jude Shavlik. "Transfer learning." *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*. IGI global, 2010. 242-264.

Vanschoren, Joaquin. "Meta-learning." *Automated machine learning: methods, systems, challenges* (2019): 35-61.

Vihvelin, Kadri. 2004. 'Free Will Demystified: A Dispositional Account', *Philosophical Topics*, 32: 427-50.

Weizenbaum J. (1966). ELIZA—a computer program for the study of natural language communication between man and machine. Communications of the ACM, 9(1), 36-45.

Zheng, L., Noroozi, V., & Yu, P. S. (2018). Joint deep modeling of users and items using reviews for recommendation. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (pp. 425-434).