8

Powerful Qualities, Phenomenal Properties, and AI

Ashley Coates

1. AI, Phenomenal Properties, and Dispositionalism

While a wide variety of approaches and techniques have been developed for producing artificial intelligences or agents, they all involve the development of algorithmic systems for moving step-by-step from certain inputs to certain outputs.[1] Physical systems are then designed that allow the algorithmic steps from an input to an output to be instantiated as dispositions to move from one physical state to another. The resulting complex dispositional systems underlie and determine the observable, behavioral dispositions of artificial agents. So, for instance, Deep Blue's disposition to make appropriate chess moves in response to prior chess moves, or Watson's disposition to correctly answer *Jeopardy!* questions, is determined by a highly complex underlying dispositional structure, which is the physical instantiation of an algorithmic structure.

As the famous examples just cited demonstrate, artificial agents of this sort can undoubtedly share certain complex behavioral dispositions with human agents. It remains an open question, though, whether an artificial agent could share enough human behavioral dispositions, and exercise them in such a way, as to be behaviorally indistinguishable from a human agent. It also remains an open question whether an artificial agent might be not only behaviorally but also mentally indistinguishable from human agents.

Here I will use "weak AI" and "strong AI" as labels for affirmative answers to these two questions:[2]

> *Weak AI* It is metaphysically possible for an algorithmic artificial agent to possess all actual human behavioral dispositions.
> *Strong AI* It is metaphysically possible for an algorithmic artificial agent to possess all actual human mental properties.

So, according to *Weak AI*, it is metaphysically possible for some artificial agent to possess a full suite of normal human dispositions toward publicly observable behavior. Such an agent would, for instance, be disposed to carry on a conversation in the same way as an ordinary human agent, as well as to take normal human actions in response to environmental stimuli. *Strong AI*, on the other hand, is the stronger claim that it is metaphysically possible for some artificial agent to share all human mental states, such as beliefs, desires, and conscious experiences.

As metaphysical claims about what is metaphysically possible for artificial agents, these claims should not be mistaken for predictions about what AI research and engineering will actually achieve. Even if *Strong AI* is true, we may actually lack the time, intelligence or resources to produce artificial agents with the relevant properties. Nonetheless, these claims are still claims about the sorts of artificial agents produced by *actual* artificial intelligence research. That is, they concern *algorithmic* artificial agents whose behavioral dispositions are determined by complex underlying dispositional structures governed by algorithms. So, any metaphysically possible worlds in which radically different sorts of artificial agents share our behavioral or mental properties and dispositions are not relevant to *Weak AI* or *Strong AI*. The two claims, then, concern what is metaphysically possible for algorithmic artificial agents, of the general sort produced by actual artificial intelligence research, irrespective of which artificial agents creatures like us, in contexts like ours, will, or even could, produce.

Strong AI has significant implications for the natures of human mental properties. As outlined above, the behavioral dispositions of artificial agents are determined just by underlying dispositional systems. So, it seems that any artificial mental properties involved in determining and explaining artificial agents' behavioral dispositions must, themselves, be dispositional. *Strong AI*, though, entails that some possible artificial agent is mentally indistinguishable from actual human agents. The apparent consequence is that any mental properties involved in determining and explaining human behavioral dispositions must be dispositional.[3]

This result, though, leads to a serious tension between *Strong AI* and an intuitive view of phenomenal properties. It is highly intuitive that *what it is to be* a phenomenal property, such as *being in pain*, consists in *what it is like* to have that property and not in causal or dispositional facts about bearers of the property. Even though *being in pain* is standardly associated with certain dispositions, pain appears to be defined by its qualitative feel rather than by those dispositions. It is also highly intuitive, though, that phenomenal properties are sometimes involved in determining our behavioral dispositions. When I am in pain, for instance, I not

only have a set of dispositions to end or mitigate that pain, but I have those dispositions, at least in part, *because* I am in pain.

So, phenomenal properties, such as *being in pain*, intuitively both have qualitative non-dispositional essences and are involved in determining and explaining human behavioral dispositions. I will call this view of phenomenal properties "*The Intuitive View*." The difficulty, then, is that, while *Strong AI* appears to entail that all mental properties involved in determining and explaining human behavioral dispositions have dispositional natures, *The Intuitive View* entails that some of those properties are non-dispositional.[4]

It is, of course, open to proponents of *Strong AI* to deny either conjunct of *The Intuitive View*. That is, they can argue that phenomenal properties, in fact, have dispositional natures,[5] or they can deny that phenomenal properties play any role in determining our behavioral dispositions. They could also attempt to square *Strong AI* with *The Intuitive View*. For instance, they could argue that an appropriately programmed artificial agent would necessarily realize qualitative phenomenal properties that feature in explanations of behavioral dispositions.

My goal here, though, is not to evaluate the prospects of these sorts of moves or directly to consider the implications of *The Intuitive View* for *Strong AI*. Instead, my aim is to clarify how a general dispositionalist approach to the metaphysics of properties bears on and informs this issue. An immediately appealing thought in this regard is that dispositionalism is inconsistent with *The Intuitive View*, as it rules out the possibility of causally relevant, non-dispositional properties. The result would be that, to the degree that one finds dispositionalism plausible, *The Intuitive View* and any difficulties that it raises for *Strong AI* are undermined. On the other hand, of course, to the degree that one is partial to the *Intuitive View*, dispositionalism would be undermined.

I am going to argue, though, that the connections between dispositionalism, *Strong AI* and *The Intuitive View* are more complicated than this initial thought might indicate. While I accept that orthodox dispositionalism about macro-properties is inconsistent with *The Intuitive View*, I also argue that dispositionalism can be squared with *The Intuitive View* by adopting a version of the "grounding theory of powerful qualities."[6] I argue further that doing so leads to an overall picture of the mind that diverges radically from the picture behind *Strong AI*, and also raises potential difficulties for *Weak AI*. The result is that dispositionalism stands in highly significant connections with the metaphysics of the dispositions of artificial agents, but these connections are more complex than they might first appear.

I begin in the next section by clarifying how I understand "dispositionalism" and arguing that orthodox dispositionalism, when applied to macro-properties,

is inconsistent with *The Intuitive View*. In Section 3, I introduce my favored account of the grounding theory of powerful qualities and apply it to phenomenal properties. In Section 4, I argue that this account of phenomenal properties ought to be combined with a libertarian, power-theoretic account of the will. In Section 5, I outline how the resulting view accommodates *The Intuitive View* and generates a picture of human cognition that is at odds with *Strong AI* and potentially problematic for *Weak AI*. The main result is that different dispositionalist views have highly significant but radically different implications for the metaphysics of the dispositions of artificial agents.

2. Orthodox Dispositionalism and Phenomenal Properties

I refer to sparse properties that, in themselves or by their natures, make a difference to the causal-modal facts about their bearers as "powers." Dispositionalism can, then, be defined as the view that there are powers. Categoricalism, on the other hand, is the view that all sparse properties are categorical properties that are, in themselves, causally and modally inert. Categorical properties have their causal-modal implications "imposed" on them by something external, such as the laws of nature.

The difference between the views can be illustrated by the common example of charge. Dispositionalists generally hold that charge is a power that, in itself or by its nature, determines charged objects' dispositions to attract or repel other charged objects. For the categoricalist, on the other hand, charge is a categorical property that, in itself, has no non-trivial causal-modal implications. Instead, charge bestows its characteristic attractive and repulsive dispositions only in conjunction with Coulomb's law. Given different laws, charge would bestow different dispositions. The dispositionalist and the categoricalist, then, do not differ over the existence of dispositions, or over which dispositions there are, but rather over whether these dispositions should be accounted for in terms of the powers of objects or in terms of something like the laws of nature.

Some dispositionalists restrict the claim that there are powers to low-level micro-properties, such as charge or mass (Bird 2007; 2016), while others extend it also to higher-level macro-properties (Ellis 2002; Molnar 2003; Mumford and Anjum 2011). As the former views have no clear bearing on the natures of higher-level artificial or mental properties, my focus will be on the latter views.[7] While I am not going to take a position here on how exactly sparse properties are to be understood, I will assume that mental properties are sparse in the relevant

sense. This assumption fits well with Schaffer's (2004) influential scientific conception of sparse properties, which Mumford (2021: §3) has, in effect, recently endorsed as the right way to identify macro-powers.

Given that there are macro-powers, though, dispositionalists still divide into those who embrace "pandispositionalism" (Mumford and Anjum 2011) and those who embrace "the mixed view" (Ellis 2002; Molnar 2003). Pandispositionalism is the view that *all* sparse properties are powers, while the mixed view, as the name implies, allows that there are both powers and sparse categorical or qualitative properties. In what follows, I will discuss both pandispositionalism and the mixed view.

Dispositionalists differ not only over which properties are powers but also over the natures of powers. On the orthodox "dispositional essentialist" approach, properties have their causal-modal implications just because they have purely dispositional *essences* or *identities* (Shoemaker 1998; Ellis 2001; Molnar 2003; Mumford 2004; Bird 2007). That is, their essences or identities are exhausted by, or are fully determined by, the dispositions or clusters of dispositions that they bestow on bearers. For instance, the essence of charge consists of the sorts of dispositions that follow from Coulomb's law. So, charged objects have these dispositions, just because it is essential to charge that its bearers have them.

The conjunction of pandispositionalism and dispositional essentialism entails that all sparse properties have exclusively dispositional essences. Given the assumption that mental properties are sparse, this view is straightforwardly inconsistent with phenomenal properties having qualitative non-dispositional essences. The result is that the conjunction of pandispositionalism and dispositional essentialism is straightforwardly incompatible with *The Intuitive View*.

The mixed view, on the other hand, allows that some sparse properties are categorical or qualitative and, so, is not in this way straightforwardly incompatible with *The Intuitive View*. I think, though, that orthodox versions of the mixed view still turn out to be incompatible with *The Intuitive View*.

The mixed view is motivated primarily by the idea that structural or spatiotemporal properties are categorical, with proponents generally holding that other properties are powers (Ellis 2002: ch. 4; Molnar 2003: ch. 10).[8] Proponents also standardly maintain the orthodox dispositional essentialist account of powers.. So, they accept that powers have exclusively dispositional essences and that the dispositions of objects are determined and explained by these sorts of powers. Categorical properties gain causal relevance just by featuring in the stimuli or manifestations of the resulting dispositions. So, while the mixed view allows that categorical properties exist and can be causally relevant, it maintains

that dispositions are fully metaphysically explained by essentially dispositional properties.

To see how this is supposed to work, we can return to the example of charge. On orthodox versions of the mixed view, the purely dispositional nature of *charge* still determines and explains the dispositions for force-exertion that follow from Coulomb's law. The nature of charge, though, also entails that the distance between two charged objects is a stimulus condition for those dispositions. So, as a categorical property, distance is, in itself, causally and modally inert, but it gains causal relevance from the nature of charge by featuring in that nature as a stimulus condition for a disposition.

Recall, though, that *The Intuitive View* holds both that phenomenal properties have qualitative non-dispositional essences and that they are involved in determining and explaining behavioral dispositions. For instance, it holds both that the essence of pain consists just in its qualitative feel and that I am disposed to take a painkiller at least partly *because* I am in pain. We have just seen, though, that the mixed view maintains that the existence of any disposition is fully determined and explained by essentially dispositional properties. So, while the mixed view is compatible with the existence of non-dispositional properties, it is incompatible with *The Intuitive View*'s implication that purely qualitative properties are involved in determining and explaining certain dispositions.

The upshot is that applying orthodox versions of dispositionalism—that is, dispositional essentialism together with either pandispositionalism or standard versions of the mixed view—to macro-properties is inconsistent with *The Intuitive View*. This result is good news for Strong AI, as it means that a major, independently motivated approach to the metaphysics of properties blocks any difficulties that *The Intuitive View* raises for *Strong AI*. On the other hand, the result looks less good for orthodox dispositionalism, as, other things being equal, it would seem best for the dispositionalist to avoid controversial commitments concerning phenomenal properties.

3. The Grounding Theory and Phenomenal Properties

In the rest of this paper, I want to consider whether, and how, the dispositionalist might accommodate *The Intuitive View*. One possibility would be simply to adopt a non-standard version of the mixed view that allows both that phenomenal properties are non-dispositional, categorical properties and that they are involved in fixing the dispositions of their bearers. This move, though, would involve a

significant retreat from the important dispositionalist idea that the dispositions of objects are fixed by their powers. Indeed, it would amount to accepting categoricalism about phenomenal properties and would be dispositionalist only in the sense that it maintains dispositionalism about other properties. As a result, the approach would also involve treating phenomenal properties and the associated behavioral dispositions in a seemingly ad hoc way.

A more promising alternative would be to turn to the "powerful qualities" view, on which properties are, in some sense, both qualitative and powerful (Martin and Heil 1999; Heil 2003; Martin 2008). This approach promises to provide a genuinely dispositionalist account of phenomenal properties, as well as to accommodate the way that, on *The Intuitive View*, phenomenal properties determine dispositions of their bearers despite being essentially qualitative. The approach also does not seem to require treating phenomenal properties in a special, potentially ad hoc, way.

An immediate difficulty with this move is that it has proven difficult to give a coherent, plausible account of how properties can be both qualitative and powerful.[9] I think though that an approach to the metaphysics of properties and dispositions, sometimes called the "grounding theory of powers," that has recently received quite a bit of attention provides a promising way around this difficulty.[10] Employing the grounding theory to make sense of powerful qualities yields a view on which properties are powerful because they ground dispositions, while they are qualitative because they have purely qualitative essences. I refer to this view as the "grounding theory of powerful qualities" (GPQ).

GPQ, though, needs to be carefully formulated if it is to pick out a genuinely power-theoretic or dispositionalist position, as orthodox versions of categoricalism seem consistent with purely qualitative properties grounding dispositions of their bearers.[11] For instance, the orthodox categoricalist could claim that a positively charged particle's having the disposition to repel positively charged particles is jointly grounded in the particle's being positively charged and the obtaining of Coulomb's law. As charge would be qualitative on the categoricalist's account, the result is that a qualitative property of an object is involved in grounding a disposition of the object.

On this categoricalist approach, though, the property is only a *partial* ground for the disposition. The disposition is *fully* grounded only in the conjunction of the property *and the law*. So, one way to distinguish GPQ from the categoricalist view would be to say that a property is powerful just if a particular's having that property *fully* grounds its having some disposition (Azzano 2021: 2967). At face value, this seems to capture the key dispositionalist idea that the property has

causal-modal significance *in itself* rather than having that significance imposed on it by a law.

As Azzano (2021: 2967) notes, though, this view seems implausible when applied to macro-properties. While it seems relatively plausible that micro-properties like charge can fully ground dispositions of their bearers, higher-level dispositions, like a billiard ball's disposition to roll down an inclined plane, generally seem to be grounded in multiple properties. The grounds for the ball's disposition to roll, for instance, seem to include the ball's having mass, being rigid and being spherical.

An alternative formulation of GPQ, taken from Tugby's formulation of the grounding theory of powers, gets around this difficulty. On Tugby's (2021: 11195) formulation, GPQ would be the view that any disposition of a particular is fully grounded in that particular's qualities. So, Tugby's view, unlike Azzano's, allows that any particular disposition of an object is grounded in multiple qualities of the object.

Tugby's view, though, does not ultimately seem to capture the spirit of dispositionalism.[12] To see the issue, consider a view on which grounding facts are metaphysically explained and determined by sui generis metaphysical laws. On this view, the real causal-modal work seems to be done by the metaphysical law, which imposes causal-modal significance on the property, rather than by the property itself. Indeed, on this view, a property could be a "thin quiddity" with no substantial nature whatsoever and still ground dispositions because a metaphysical law links it to those dispositions.

The lesson here, I think, is that facts about "meta-grounding" are relevant to whether a proposed formulation of GPQ is genuinely dispositionalist. What matters is not just how properties ground dispositions but also *how that grounding fact is metaphysically explained*. In particular, it matters whether the property's capacity to ground dispositions is imposed on it by something external, like a law of some sort, or stems from the property itself or from the property's own nature.

Inspired by this idea, I have elsewhere (Coates 2023) proposed the following account:

> The meta-grounding theory (MGT): A qualitative property, F, is powerful iff (a) at least some instances of F at least partially ground dispositions of their bearers and (b) the essence of F, at least partially, grounds (a).

The core idea behind this account is that the meta-grounding claim in (b) captures the idea that powerful properties contribute to the causal-modal facts, *just because of their own natures and not because that contribution is imposed on*

them by anything like a law of nature or a law of metaphysics. Given this idea, powerful qualities have purely qualitative essences that, nonetheless, determine and explain how their instances enter into the grounds of dispositions.

To see how this works, we can return to the example of a billiard ball's disposition to roll. While the essence of *being spherical* appears to consist just in its qualitative geometrical definition, that essence still *explains* how the ball's being spherical is involved in determining and explaining the ball's disposition to roll. This point plausibly explains why shape properties are such common examples of apparent powerful qualities. While these properties appear to have non-dispositional, qualitative essences, those essences still seem capable of explaining how shape properties help determine the dispositions of their bearers (Lowe 2010: 20–1; Yates 2018: 4538).

Importantly, though, MGT seems to apply to phenomenal properties in much the same way that it applies to shape properties.[13] We have already seen that phenomenal properties intuitively have qualitative essences that consist in *what it is like* to have the property. For instance, what it is *to be in pain* plausibly consists in the subjective painfulness of pain. Nonetheless, S's being in pain appears, at least partially, to ground certain behavioral dispositions of S, such as the disposition to remove the painful stimulus or to mitigate the pain experience. Moreover, it is also strongly intuitive that this grounding connection holds at least partly because of the qualitative phenomenal essence of pain. Intuitively, it is because of how pain feels that S's being in pain at least partially grounds S's having the relevant sorts of behavioral dispositions.

I need to emphasize that, according to MGT, the essence of a powerful quality is *purely* qualitative and does not include facts or sentences about the dispositions that the quality grounds. The essence of a shape quality consists *exclusively* of its geometrical definition, while the essence of a phenomenal property consists *exclusively* of *what it is like* to have the property. The key idea is that these purely qualitative essences still explain why certain instances of the properties ground certain dispositions. As Lowe (2010: 20–1) and Yates (2018: 4538) emphasize, the geometrical essence of being spherical does not include anything about rolling but, nonetheless, perfectly well explains the disposition of certain spheres to roll. Similarly, while the phenomenal nature of pain is not defined in terms of any dispositions, that phenomenal nature intuitively partially explains the aversive dispositions that come with pain.

On MGT, then, powerful qualities do not have even partially dispositional essences. Instead, qualities are powerful, on this view, because their *purely qualitative* essences have ultimate metaphysical responsibility for the dispositions

that they bestow on bearers. In line with my earlier definition of powers, MGT holds that qualities are powerful because they, by their natures, make a difference to the causal-modal facts about their bearers, even though they do not have dispositional essences.

The resulting account of phenomenal properties and their causal-modal implications is closely related to the "phenomenal powers" view developed by Hedda Hassel Mørch. Unlike the account just given, though, Mørch's (2018: 304; 2020: 1082) view requires that an instance of a phenomenal property *fully* grounds the relevant dispositions. In support of this commitment, Mørch (2019; 2020) argues that, at least given that pain has causal-modal consequences in virtue of its nature, it is not genuinely conceivable that I can be in pain without having pain-aversive dispositions.

Philip Goff (2020: 1092), though, has recently argued against Mørch that *irrational* agents, who are disposed to pursue pain and avoid pleasure, seem perfectly conceivable. Given that conceivability entails metaphysical possibility, though, the result is that *being in pain* does not necessitate having pain-aversive dispositions. So, given the orthodox view that the full grounds for a state metaphysically necessitate that state, *being in pain* cannot fully ground the aversive dispositions that I have when I am in pain.

This point seems right. Although creatures who were "wired up" to pursue pain and avoid pleasure might be irrational, and would certainly be unfortunate, I do not think that they are, in any clear way, *impossible*. Unlike Mørch's view, though, the account of phenomenal properties just given is compatible with this result. While Mørch requires that a phenomenal power fully grounds some disposition of its bearer, MGT allows that powerful qualities only partially ground dispositions. The proponent of MGT can, then, claim that having a phenomenal property only fully grounds behavioral dispositions in conjunction with *being rational*.[14]

4. The Role of Rationality and the Will

I think, though, that there is still a difficulty in this vicinity for MGT and the associated account of phenomenal properties. A seemingly plausible way to flesh out the grounding of pain-aversive dispositions in the conjunction of *being in pain* and *being rational* is as follows:

(1) S's having pain-aversive disposition, D, *is grounded in* S's being rational & S's being in pain
is grounded in

> Pain having its phenomenal essence & the essence of being rational consisting partly in having D when having a property with the phenomenal essence of pain.

The thought here is that pain-aversive dispositions are grounded in *being rational* and *being in pain*, because having these dispositions in response to the feel of pain is constitutive of what it is to be rational in the relevant sense. While MGT entails that this sort of account suffices for *being in pain* to be a powerful quality, this result does not actually seem right.

To see the problem, assume that Q in the following fact is a thin quiddity that lacks any substantial qualitative nature:

> (2) *x*'s having disposition, D, *is grounded in x*'s being R & *x*'s being Q
> *is grounded in*
> Q having essence φ & the essence of being R consisting partly in having D when having a property with essence φ.

Given MGT, (2) entails that Q is a powerful quality in exactly the same way that (1) entails that *being in pain* is a powerful quality, but this is surely the wrong result. Q has no substantial nature nor does it have any causal-modal significance in itself or by its own nature. Instead, its causal-modal significance is entirely imposed on it by the essence of R and, in particular, the manner in which the intrinsically inert essence of Q features in the essence of R. So, although an instance of Q plays a role in grounding a disposition partly because of the essence of Q, Q still has an intrinsically inert nature that derives its causal modal significance from the essence of R. Consequently, (2) ought not to suffice for Q to be a powerful quality.

In the same way, (1) ought not to suffice for *being in pain* to be a powerful quality, because the causal-modal work is really being done by the nature of *being rational*, which "imposes" a dispositional character on *being in pain*. Consequently, (1) leaves it possible that *being in pain* is, in itself, a causally and modally inert property, and should not suffice for *being in pain* to be a powerful quality.

I think these sorts of cases indicate that MGT needs to be modified to ensure that the relevant properties are genuinely distinct from each other and make distinct contributions to the dispositions of things. To meet this requirement, I will add a third condition to MGT:

> MGT*: A qualitative property, F, is powerful iff (a) at least some instances of F at least partially ground dispositions of their bearers, (b) the essence of F at least partially grounds (a), and (c) F is genuinely distinct from any other properties involved in grounding the disposition in (a).

I understand "genuinely distinct" here to mean that neither the property or its essence features in the other property's essence. This additional condition ensures that a powerful property (or its essence) does not gain causal-modal significance just by entering into the essence of some other property. In particular, it rules out the possibility that a phenomenal property like pain could count as a powerful property just by the way it enters into the essence of being rational. Instead, in line with the core dispositionalist idea, the property's essence must have causal-modal significance "in its own right." This additional condition also fits with paradigmatic cases, such as the case of the billiard ball. The ball's being spherical is a partial ground for its disposition to roll, because of its own distinctive geometrical nature, and not because of any essential connection between *being spherical* or its essence and some other ground for that disposition.

In the present context, though, this distinctness condition constitutes a highly significant constraint. It requires that *being in pain* and *being rational* jointly ground certain dispositions, *even though no connection between pain, or the phenomenal character of pain, and those dispositions is definitive of being rational*. So, *being in pain* and *being rational* must, so to speak, "generate" a novel connection between pain and certain dispositions without any such connection being inscribed into the nature of *being rational*.

This requirement is significant, at least in part, because it rules out a broadly functionalist account of rationality. On such an account, *being rational* is constituted by particular rational dispositions. To be rational is just to have a set of dispositions to act in the right way, or to form the right intentions or volitions, in response to the right sorts of stimuli. For instance, to be rational is constituted, in part, by having pain-aversive dispositions in response to pain or to the painfulness of pain. Consequently, (c) rules out the possibility of invoking MGT* to give a dispositionalist account of pain that grounds behavioral dispositions in the conjunction of *being in pain* and this sort of functionalist rationality.

What is required, instead, is a notion of *being rational* that is defined independently of specific dispositions to particular actions in response to particular stimuli. Certain power-theoretic, libertarian accounts of the will, though, involve just such a notion. I will focus on E. J. Lowe's account of the will,[15] although, as I will explain shortly, not all of Lowe's commitments seem essential in the present context.

On Lowe's view, the will is a non-causal, open-ended, irreducible "rational power" to form volitions "in the light of reasons." Lowe holds that *nothing* causes the volitions that are the result of the exercise of the will, or even the probabilities of those volitions. Instead, volitions are explained by an agent's exercise of distinctively *rational powers* in the light of reasons. In other words, volitions

have a rational, non-causal explanation, in terms of agents' reasons and their rational power to form volitions in response to reasons. Lowe also holds an externalist view of reasons on which reasons are objective states of affairs that exist independently of an agent's beliefs, desires or other mental states.

So, for Lowe, the sort of rationality involved in being an agent consists in having an open-ended, non-causal capacity to recognize and act on objective reasons for action. Most important in the present context is that this capacity is not constituted by dispositions for particular volitions or actions in response to specific reasons. Instead, it consists in a *general*, *open-ended* capacity to recognize and act *for* independently existing reasons.

Given this conception of *being rational*, the grounding of pain-aversive dispositions in the conjunction of *being in pain* and *being rational* could be grounded as follows:

> (3) S's having a pain-aversive disposition *is grounded in* S's being rational & S's being in pain
> *is grounded in*
> Pain having its phenomenal essence & the essence of being rational consisting in having a general, open-ended capacity to act for reasons.

(3) suggests an intuitive account of how *being in pain* enters into the grounds of behavioral dispositions. On this account, *being in pain* constitutes an objective reason for pain-aversive actions, just because of *what it is like* to be in pain. Put differently, the objective, inherent badness of *what it is like* to be in pain metaphysically suffices for pain to constitute a reason for such actions. *Being rational*, in turn, consists in having a *general*, *open-ended* capacity to recognize reasons and form volitions based on them. Consequently, *being in pain* and *being rational* jointly ground pain-aversive dispositions, because the essences of these properties ensure that anyone who has both of them both has a reason for pain-aversive actions and is responsive to such reasons.

(3), unlike (1), also entails that *being in pain* satisfies MGT*. The key point in this respect is that in (3), unlike in (1), *being rational* is defined in terms of a general rational power, and not in terms of any particular dispositions involving pain or the essence of pain. Consequently, *being in pain* and *being rational* are genuinely distinct properties, and the phenomenal nature of pain, via its intrinsic badness, plays its own distinctive role in determining the causal-modal consequences of *being in pain*. Lowe's account of the will and of rationality, then, allows us to use MGT* to give an intuitive, dispositionalist account of how *being in pain* enters the grounds of pain-aversive dispositions.

As I indicated above, though, not all aspects of Lowe's account seem necessary here. In particular, Lowe's highly controversial claims that the will is an entirely "spontaneous" power—one that is not probablistically structured—and that it is a non-causal power do not seem necessary. What seems essential, instead, is that the will is an *open-ended, general* power to act for reasons that, together with the existence of certain reasons, can ground specific dispositions toward, or powers for, intentions, volitions or actions.[16]

Something like Lowe's externalism about reasons also seems necessary. It seems possible, after all, that creatures that are "wired-up wrong" might experience pain, but, nonetheless, form desires to pursue pain. So, if pain states count as reasons only via their connection to an agent's actual desires and motivational states, then *being in pain* and *being rational* do not seem capable, in themselves, of grounding any behavioral dispositions. Instead, what is required is a view on which the intrinsic, objective badness of *being in pain* is sufficient to make it a reason for pain-aversive actions, irrespective of a creature's actual desires and behavioral states.

I think we now have in place a genuinely dispositionalist account of *being in pain* that also satisfies *The Intuitive View*. The account is dispositionalist, because it entails that *being in pain*, by its nature and in its own right, makes a distinctive contribution to the causal-modal facts. The account, furthermore, provides a fully dispositionalist grounding of the relevant dispositions by grounding them just in the conjunction of *being in pain* and a dispositionalist account of the will. The account, though, also directly entails that *The Intuitive View* holds of *being in pain*, as it entails that *being in pain* both has a purely qualitative essence and is involved in determining and explaining pain-aversive dispositions.

The account, of course, comes with controversial commitments, as it depends on Lowe's version of externalism about reasons and something like his account of the will as a general, open-ended power to respond to reasons. My goal here, though, is not to defend these commitments, but rather to show that assuming them allows for an account of pain that is both dispositionalist and in line with the *Intuitive View*. Doing so is sufficient to fulfil my goal of showing that it is possible to reconcile these two views and of clarifying what such a reconciliation might look like.

5. Implications for the Metaphysics of AI

Moreover, extending this account to other phenomenal properties that have clear motivational or normative content seems straightforward. An obvious

example is *experiencing pleasure*, but other possible examples include *having a desire* or experiencing specific emotions.[17] At least initially, it seems plausible that these properties all have a qualitative phenomenal nature that can explain how they, along with *being rational* in the appropriate sense, can ground dispositions toward certain actions.

It may also be possible, though, to extend the view to phenomenal properties that do not have obvious normative or motivational content. To do so, one could begin by claiming that reasons for beliefs are, in general, grounded in phenomenal states.[18] On this view, for instance, I might claim that certain of my phenomenal states, at least partially, ground my belief that the wall in front of me is yellow. Given this view, phenomenal properties that lack obvious motivational or normative significance could still, in conjunction with my rational powers, ground my behavioral dispositions by grounding my reasons for beliefs. To use a simple example, certain of my phenomenal states will be involved in grounding my disposition to agree that the wall in front of me is yellow via grounding my belief that it is yellow.

This very brief sketch obviously only points towards a possible extension of the view discussed here. I hope it does enough, however, to motivate taking seriously the possibility that this view may provide a way to reconcile dispositionalism with a general endorsement of *The Intuitive View*.

If anything like the view just sketched were true, though, then human behavioral dispositions and the behavioral dispositions of algorithmic artificial agents would be fixed in fundamentally different ways. On the view sketched here, human behavioral dispositions would be largely determined by the interaction of open-ended rational powers and essentially qualitative phenomenal states. In algorithmic artificial agents, on the other hand, actions are determined by complex dispositional systems to move from certain stimuli to certain actions. While these dispositional structures vary in many respects, they are always broadly functionalist structures that are constituted by dispositions to move from specific stimuli to specific manifestations.

Note that this point holds also for machine learning systems, even though the behavioral dispositions or outputs of those systems are not initially programmed into them. As Patterson and Gibson (2017: 1) summarize it, "Fundamentally, *machine learning* is using algorithms to extract information from raw data and represent it in some type of *model*. We use this model to infer things about other data we have not yet modelled." The ultimate outcomes of this sort of system are determined by a generally highly complex process in which models are constructed from data and then used to generate outputs in response to novel

inputs. So, the system is not initially programmed to produce specific outputs, and its actual outputs often cannot be predicted.

Nonetheless, any particular output is still the result of a system of discrete dispositions governed by a set of algorithms. The system's algorithms ensure that it is disposed to generate certain models in response to specific data and that, given certain models and inputs, it is disposed to produce particular outputs. So, *given the initial inputs*, a complex system of underlying discrete dispositions to move from particular inputs to certain outputs determines the ultimate output of the system. The complexity of this dispositional system, and the manner in which certain of the dispositions depend on others, makes the outputs unpredictable. However, it remains true that the actual outputs are ultimately determined by a system of specific dispositions.

On the other hand, the rational power of the will is not defined or determined by any dispositional system linking specific inputs and outputs. Instead, this general, rational power together with a particular set of reasons, potentially constituted largely or entirely by phenomenal states, is itself the ultimate grounds for any dispositions to act in specific ways in response to particular reasons. So, on the view developed here, behavioral dispositions would be determined by a general, open-ended power of the will—defined independently of any set of discrete dispositions—working in conjunction with an agent's phenomenal states. This picture differs sharply from the sort of highly complex, dispositional system involved in determining the behavioral dispositions of algorithmic artificial agents, even given that those agents employ machine learning.

The view developed here, then, leads to a picture of human cognition that departs radically from the picture that follows from *Strong AI*. The core difference is a difference in the sorts of powers or dispositions that underlie and determine behavioral dispositions in the two cases. In the former case, these dispositions are determined by open-ended, general, irreducibly normative powers working together with powerful phenomenal qualities. In the latter case, they are determined by underlying algorithmic dispositions to move step-by-step from particular inputs to particular outputs. So, although I set out to develop a view that, just by respecting *The Intuitive View*, would be in tension with *Strong AI*, the resulting view looks far more deeply and fundamentally incompatible with *Strong AI*'s overall picture of the human mind.

My primary focus here has been on *Strong AI*, but before concluding I want to briefly return to *Weak AI*. Recall that *Weak AI*, as I defined it, is the view that it is metaphysically possible for an algorithmic artificial agent to possess a full suite of human behavioral dispositions and, so, be behaviorally indistinguishable

from a human agent. The behavioral dispositions of any such artificial agent would be determined by elaborate dispositional or causal systems moving stepwise from specific inputs to specific outputs. Such systems, though, are highly computationally demanding and, as a consequence, are also spatially or temporally demanding. Each of the computational and causal steps that must be taken to move from a stimulus to a particular output takes up some available space or time, with the two having to be traded off against each other.

On the view of human cognition sketched here, on the other hand, specific dispositions could very often be grounded in the conjunction of a qualitative state and a general, standing rational power. The apparent consequence is that the underlying structure required for cognition and behavior could be greatly simplified. This structure need only implement the qualitative states and the general, standing power in order to ground the specific dispositions. The result is a potentially far less computationally, spatially and temporally demanding way of moving from stimuli, such as sensory stimuli, to behavioral outputs.

There is, of course, a substantial question about how this kind of cognitive structure would connect to the complex dispositional structure of the brain. While I cannot address this issue in any detail here, both Lowe and O'Connor, who I cited as proponents of the conception of the will that I am working with, are emergentists about mental properties and powers. While they accept that these properties and powers causally depend on the brain, they deny that they are constituted by brain states. Consequently, on this view, the dispositional complexity of the brain states responsible for exercises of the will need not imply any corresponding complexity in the will or in phenomenal properties. So, it remains possible that the latter are simple and that, by grounding specific behavioral dispositions, they simplify the dispositional structure required for cognition and behavior.

If this is right, then any algorithmic artificial agent faces the challenge of using algorithmic systems to simulate the behavioral dispositions of far more efficient qualitative and normative systems. While there is certainly no proof here that doing so is not possible, it is also not obvious that it is. Minimally, if human cognition works in anything like this way, it raises an in-principle question about Weak AI that simply does not arise if human cognition is fundamentally algorithmic.

6. Conclusion

My goal here has not been to argue for the position developed over the last three sections. Instead, recall that my goal in these sections was to consider whether

dispositionalism, when applied to macro-properties, can be reconciled with *The Intuitive View* and, if so, what such a reconciliation might look like. I have argued that this reconciliation can be achieved by combining a version of the grounding theory of powerful qualities, MGT*, with existing dispositionalist, libertarian accounts of the will. I have also argued that the resulting view entails that human cognition works radically differently to artificial cognition. As a result, the view is seemingly incompatible with *Strong AI* and also leads to questions about *Weak AI*.

Note also that the view developed here is not a new version of dispositionalism constructed just to accommodate *The Intuitive View*. Instead, the position is built out of existing, independently motivated dispositionalist positions. Consequently, the key result is that the combination of certain existing dispositionalist views can generate a position that is hospitable to *The Intuitive View*, inhospitable to *Strong AI* and in some tension with *Weak AI*.

This result contrasts sharply with the results that I earlier reached about the application of orthodox dispositionalism—the combination of dispositional essentialism with either pandispositionalism or the mixed view—to macro-properties. I argued that this sort of dispositionalism is inconsistent with *The Intuitive View* and, so, is amenable to *Strong AI*. While I did not directly consider this sort of dispositionalism's implications for *Weak AI*, it would not seem to raise any difficulties for *Weak AI*.

My main conclusion, then, is that, while dispositionalism, as such, does not have specific implications for the metaphysics of AI and of phenomenal properties, different kinds of dispositionalism have highly significant and radically different implications in these areas. In addition to defending this point, I have attempted to map out some of these important connections. I hope that doing so provides the basis for further work both clarifying how dispositionalism bears on the metaphysics of AI, and determining which conclusions we ought ultimately to draw from these connections.[19]

Notes

1 See Bringsjord and Govindarajulu (2022) for a helpful overview.

2 These labels go back to Searle (1980), although the way I use them here is somewhat different and is essentially the same as Bringsjord and Govindarajulu (2022) use of the terms.

3 This point is closely related to the well-known point that *Strong AI* appears to lead to a functionalist account of the mind. This connection has been widely discussed in

connection with Searle's (1980) hugely influential "Chinese Room Argument." For an overview, see Cole (2020).

4 The underlying difficulty, of course, is the well-known problem of squaring an intuitive view of phenomenal properties, and associated "qualia," with a functionalist account of the mind. For an overview see Levin (2018).

5 See Gozzano (2018) for this sort of argument. Thanks to a reviewer for suggesting this example.

6 The sort of view that I am using this label to refer to was originally introduced by Jacobs (2011) and Tugby (2012) and has recently received a considerable amount of attention (Yates 2018; Coates 2021; Azzano 2021; Tugby 2021; Kimpton-Nye 2021). Note, though, that, while some of these authors explicitly identify the approach as a version of the powerful qualities view (Jacobs 2011; Yates 2018; Coates 2021; Azzano 2021), others do not (Tugby 2021; Kimpton-Nye 2021).

7 Bird (2018) does actually allow that mental properties, along with other evolved properties, may be an exception to his restriction of dispositionalism to the fundamental level. My point, though, is just that any view on which there are, in fact, no higher-level powers would be irrelevant in this context.

8 An exception is "numerical identity of parts," which Molnar (2003: 160) also identifies as a categorical property.

9 The canonical version of the powerful qualities view is the so-called "identity theory," on which a property's powerfulness is identical to its qualitativity (Martin and Heil 1999; Heil 2003; Martin 2008). The coherence of this view, though, has often been called into question, and Taylor (2018) has recently given a sustained argument that good sense cannot be made of it. In addition to the grounding theory of powerful qualities that I discuss in the main text, other alternatives to the identity view are a "compound" view (Taylor 2018: 1438) and a "dual-aspect" view (Giannotti 2021).

10 See note 5 for references to the literature on the grounding theory. As I note there, some, but not all, proponents of the view intend it as a version of the powerful qualities view.

11 The discussion over the next couple of pages of how best to formulate GPQ is a condensed version of an argument that I have made in more detail elsewhere (reference redacted).

12 The point discussed in this paragraph is also discussed by Vetter (2021: footnote 12) and acknowledged by Tugby (2022: 23). Vetter's conclusion, in line with mine here, is that, while Tugby's view is broadly anti-Humean, it is not dispositionalist. Vetter also notes that Tugby has confirmed in personal communication that this is why he does not call his view "dispositionalist." In line with this, Tugby (2022: 23–5) argues that the view is a kind of "dispositional realism" that is distinct from categoricalism, but he does not claim that it is genuinely dispositionalist. For this reason, the point made in the main text is not a criticism of Tugby's view in itself, but rather a reason not to treat it as a version of the powerful qualities view or of dispositionalism more broadly.

13 Both Tugby (2012: 730) and I (2021: 8358) have used phenomenal properties as potential examples of GPQ, although neither pursues this connection in detail.

14 A reviewer has pointed out that, at face value, this claim seems too strong, because it seems that many animals that are not rational have pain-aversive dispositions. I think, though, that the notion of rationality developed in the next section blunts the force of this concern. This sort of rationality involves a responsiveness to the kind of objective value or disvalue involved in experiences of pleasure and pain while not necessarily requiring capacities for high-level reasoning. While more could clearly be said about this topic, I think that this conception of rationality means that it is at least not obvious that animals with pain-aversive dispositions cannot be rational in the relevant sense.

15 This account is developed in detail in the second half of Lowe (2008).

16 In these respects, Timothy O'Connor's power-theoretic account of the will also seems to do the necessary work (for a detailed and up-to-date account of his view see O'Connor 2021). Unlike Lowe, O'Connor allows both that the exercise of the will *causes* intentions and that the will is probabilistically structured, without being determined, by prior conditions such as an agent's motivational states. O'Connor, though, agrees that the will is a fundamental and open-ended power that is not defined or determined by connections between particular reasons or motivational states, on the one hand, and certain actions, volitions or intentions, on the other. Indeed, O'Connor explicitly claims that powers to form specific intentions are grounded in the general power for choice that constitutes the will, alongside facts about an agent's knowledge and motivational states.

17 In developing her phenomenal powers view, Mørch also focuses on pleasure and pain and suggests that the view can be expanded to "emotional phenomenal properties, such as anger or joy" (Mørch 2018: 303).

18 See Smithies (2019) for an extended, sophisticated argument that phenomenal consciousness is central to epistemic justification. In line with my discussion in this paragraph, Smithies (2019: 21) also points out that this conclusion means that phenomenal consciousness is essential to the rational justification for actions.

19 I'd like to thank the editors of this volume, Anna Marmodoro and Bill Bauer, and an anonymous referee for helpful comments on earlier versions of this paper.

References

Azzano, L. (2021), "Dispositionality, Categoricity, and Where to Find Them," *Synthese*, 199: 2949–76.

Bird, A. (2007), *Nature's Metaphysics: Laws and Properties*, New York: Oxford University Press.

Bird, A. (2016), "Overpowering: How the Powers Ontology Has Overreached Itself," *Mind*, 125 (498): 341–83.

Bird, A. (2018), "Fundamental Powers, Evolved Powers and Mental Powers," *Aristotelian Society Supplementary Volume*, 92 (1): 247–75.

Bringsjord, S. and N. S. Govindarajulu (2022), "Artificial Intelligence," in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2022 Edition), forthcoming https://plato.stanford.edu/archives/fall2022/entries/artificial-intelligence.

Coates, A. (2021), "Making Sense of Powerful Qualities," *Synthese*, 198: 8347–63.

Coates, A. (2023), 'The Meta-Grounding Theory of Powerful Qualities', *Philosophical Studies*. https://doi.org/10.1007/s11098-023-01982-y.

Cole, D. (2020), "The Chinese Room Argument," in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 Edition), https://plato.stanford.edu/archives/win2020/entries/chinese-room.

Ellis, B. (2002), *The Philosophy of Nature: A Guide to the New Essentialism*, Chesham: Acumen.

Goff, P. (2020), "Revelation, Consciousness+ and the Phenomenal Powers View," *Topoi*, 39: 1089–92.

Gozzano, S. (2018), "The Dispositional Nature of Phenomenal Properties," *Topoi*, 29: 1045–55.

Heil, J. (2003), *From an Ontological Point of View*, New York: Oxford University Press.

Jacobs, J. D. (2011), "Powerful Qualities, Not Pure Powers," *The Monist*, 94: 81–102.

Kimpton-Nye, S. (2021), "Reconsidering the Dispositional Essentialist Canon," *Philosophical Studies*, 178 (10): 3421–41.

Levin, J. (2021), "Functionalism," in E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2021 Edition), https://plato.stanford.edu/archives/win2021/entries/functionalism.

Lowe, E. J. (2008), *Personal Agency: The Metaphysics of Mind and Action*, New York: Oxford University Press.

Lowe, E. J. (2010), "On the Individuation of Powers," in A. Marmodoro (ed.), *The Metaphysics of Powers: Their Grounding and Their Manifestations*, 8–26, New York: Routledge.

Martin, C. B. (2008), *The Mind in Nature*, New York: Oxford University Press.

Martin, C. B. and J. Heil (1999), "The Ontological Turn," *Midwest Studies in Philosophy*, 23 (1): 34–60.

Molnar, G. (2003), *Powers: A Study in Metaphysics*, New York: Oxford University Press.

Mørch, H. H. (2018), "The Evolutionary Argument for Phenomenal Powers," *Philosophical Perspectives*, 31: 293–316.

Mørch, H. H. (2019), "Phenomenal Knowledge *Why*: The Explanatory Knowledge Argument against Physicalism," in S. Coleman (ed.), *The Knowledge Argument*, 223–53, Cambridge: Cambridge University Press.

Mørch, H. H. (2020), "Does Dispositionalism Entail Panpsychism?," *Topoi*, 39: 1073–88.

Mumford, S. (2004), *Laws in Nature*, London: Routledge.

Mumford, S. (2021), "Where the Real Power Lies: A Reply to Bird," 130: 1295–308.
Mumford, S. and R. L. Anjum (2011), *Getting Causes from Powers*, New York: Oxford University Press.
O'Connor, T. (2021), "Freewill in a Network of Powers," in W. M. R. Simpson, R. C. Koons, and J. Orr (eds.), *Neo-Aristotelian Metaphysics and the Theology of Nature*, 151–68, New York: Routledge.
Patterson, J. and A. Gibson (2017), *Deep Learning: A Practitioner's Approach*, Sebastopol: O'Reilly Media.
Schaffer, J. (2004), "Two Conceptions of Sparse Properties," *Pacific Philosophical Quarterly*, 85 (1): 92–102.
Searle, J. (1980), "Minds, Brains and Programs," *Behavioral and Brain Sciences*, 3 (3): 417–24.
Shoemaker, S. (1998), "Causal and Metaphysical Necessity," *Pacific Philosophical Quarterly*, 79 (1): 59–77.
Smithies, D. (2019), *The Epistemic Role of Consciousness*, New York: Oxford University Press.
Taylor, H. (2018), "Powerful Qualities and Pure Powers," *Philosophical Studies*, 175 (6): 1423–40.
Tugby, M. (2012), "Rescuing Dispositionalism from the Ultimate Problem: Reply to Barker and Smart," *Analysis*, 72 (4): 723–31.
Tugby, M. (2021), "Grounding Theories of Powers," *Synthese*, 198 (12): 11187–216.
Tugby, M. (2022), "Dispositional Realism without Dispositional Essences," *Synthese*, 200: 1–27.
Vetter, B. (2020), "Explanatory Dispositionalism," *Synthese*, 199: 2051–75.
Yates, D. (2018), "Inverse Functionalism and the Individuation of Powers," *Synthese*, 195: 4525–50.

Part Four

Artificial Moral Dispositions