# Is AI Capable of Aristotelian Full Moral Virtue?
# The Rational Power of *phronesis*, Machine Learning and Regularity

Ruth Groff and John Symons

*Appears in Artificial Dispositions: Investigating Ethical and Metaphysical Issues*
William A. Bauer and Anna Marmodoro (Editors) Bloomsbury

Abstract:

We argue that an artificially intelligent artifact cannot be virtuous as per Aristotle's depiction of the *phronimos* (the same is probably true in relation to a Platonic account of what it is to be moral, and perhaps others too, but for heuristic purposes we restrict our discussion to the Aristotelian case.) Crucially, the exercise of *phronesis* is neither rule-based nor generalization-based. It is, by definition, a power that cannot be transposed into specifiable stimulus-response sequences or into a program. The analysis has implications for areas of inquiry other than the philosophy of AI. Specifically, at the level of the meta-philosophy of the ontology of moral philosophy, the lesson to be learned is that defending Aristotelian full moral virtue as an account of moral agency, be it for humans or for AI, commits one to a certain kind of metaphysics. Similarly, even just affirming the existence of a power such as *phronesis* within one's metaphysics commits one to a powers ontology that does not bottom out in regular sequences.

Keywords: AI and ethics; powers; Aristotle; machine learning

Is AI Capable of Full Moral Virtue?
The Rational Power of *phronesis*, Machine Learning, and Regularity

Efforts to build artificially intelligent systems that comport with our moral values face well-known technical and theoretical challenges. In the recent development of AI, many of these challenges have been bundled together under what is known as the alignment problem. The alignment problem in AI refers to the challenge of designing intelligent systems that pursue outcomes in keeping with human values and goals. The problem arises because the objectives that AI systems optimize can diverge from human values in unexpected and potentially harmful ways, resulting in unintended consequences. The goal is to ensure that AI systems behave in a manner that is safe, transparent, and consistent with human values.

Needless to say, the alignment problem has not been solved. Nonetheless, technological innovation continues apace, and robots and AI assistants are being used in a wide range of situations, including scenarios involving high risk decisions in contexts such as warfare and medicine. Social robots, chatbots, and other AI systems are also being developed to interact with humans in a wide range of everyday situations such as customer service applications, academic writing, recommendation engines, and a wide range of other commercial and governmental applications. Given that we cannot simply start from scratch, and that we are dealing with practical challenges posed by existing technologies, striving for ethical AI requires us to connect any answer that we might give to the question: "What do we regard as the ideal, in this regard?" to facts about what we are currently capable of creating, and about what the current and near-term state is of our technology.

As a way of thinking through both the technical and moral issues at hand, we are going to consider whether an AI can be developed that approximates something like an Aristotelian *phronimos* – specifically, whether or not an AI can be a bearer of not just of 'moral virtue,' or what is

sometimes called 'natural moral virtue,' but of so-called 'full moral virtue.' If our AI systems could manifest full moral virtue, then the alignment problem would be solved.

In this paper, we have organized our discussion of the general question of ethics and AI around a particular moral theory because - and in order to highlight the fact that - different moral theories require, and thus presuppose, that different capacities be had by moral agents -- which capacities, in turn, would involve distinct kinds of technological implementations. Kantian moral agents must be capable of generating (and acting upon) the deliverances of pure practical reason. Utilitarian moral agents *a la* John Stuart Mill have to be able to calculate, and perhaps experience, both quantities and qualities of pleasure. Humean or Smithian moral agents must be able to reliably feel certain sentiments in reaction to given situations. Existentialist moral agents must have free will, and be capable of expressing it authentically. Platonists have to be such that they can be compelled by the form of the Good. Etcetera.

We have two reasons for asking if an AI could be moral as per Aristotle (rather than as per Kant or Mill or Hume or Sartre or even Plato). First, we take the Aristotelian model of moral agency to be diagnostically apposite. Why? To begin with, Aristotle's notion of full moral virtue resists a straightforwardly algorithmic treatment in ways that we believe sheds light on the ethics of AI. Accordingly, it can help us to see what the limits might be to what some philosophers have called 'artificial wisdom,' (Sullins 2020). Admittedly, Platonic and existentialist approaches would function in the same way heuristically – and arguably a sentiment-based model might do something analogous if it turns out that qualitative experiences themselves are essential to such a model and that such things cannot be replicated in an AI. But the advantage of turning to Aristotle here is that, given the contours of contemporary analytic metaphysics, the Aristotelian case is best situated to illustrate not just that different moral theories presuppose agents with different capacities, but also that different moral theories presuppose different underlying metaphysical frameworks. The tacitly

held underlying metaphysical framework, which we shall refer to as 'nomological' (and which will be addressed in Section 3, below), has been increasingly criticized (in the contemporary analytic context) by Aristotelians and neo-Aristotelians. Given the relative prominence of the Aristotelian alternative to the nomological framework, and the importance of virtue ethical perspectives in the ethics of AI, a specifically Aristotelian test of the potential limits of the moral capacities of machines seems to us to be the right choice for thinking through the alignment problem. Second, we are genuinely interested in Aristotle's account of the *phronimos* as a moral ideal Bracketing the practical matter of the sophistication of our current technology, it is worth asking if, ***in principle***, a machine could make wise judgement calls as such calls are conceptualized by Aristotle. Take, for instance, the phenomenon of autonomous weapons, such as drones or other military robots. Intelligent versions of such systems would have to be able to make decisions and navigate their environments independently of human guidance. If one is an Aristotelian, or even just attracted to aspects of Aristotle's moral theory, it makes sense to ask whether or not we can build an AI that is capable of the kind of virtuous practical judgment that Aristotle would expect from a responsible human adult (just as if one were a Kantian, say, it would make sense to ask if a machine could be programmed to act in accordance with the strictures of pure practical reason).

1. The Aristotelian Picture ~ Ethics

Inasmuch as we are looking to Aristotle's account as a criterion for whether an AI can be moral, let us begin by briefly rehearsing the position. As Aristotle has it, what is good without qualification for anything is that it exist fully as a substance or artifact of its kind. Aristotle uses the term *energeia* to denote this condition of actualization. Accordingly, the first task in the *Nicomachean Ethics* (*NE*) is to determine what kind of a substance human beings are, so as to thereby identify the activity the excellent doing of which constitutes the full being, which is to say the flourishing, of the

paradigmatic members of our kind (in Aristotle's view, these would be wealthy, intelligent, and able-bodied Greek men). Aristotle tells us that the distinctively human activity is activity "in conformity with a rational principle or, at least, not without it" [*Nicomachean Ethics* 1098a] and identifies two versions of said kind-essential activity, two versions of flourishing – both of which involve the display of rational powers. [*NE*, Book 10] The first version, which Aristotle calls 'contemplation,' is the enactment of *sophia* (itself a combination of *nous* and *epistêmê*). Contemplation is the act of grasping the forms of things (specifically, of those things – "things" as a count noun – that have them), i.e., of *nous*, and of then putting forward explanations based upon those forms, i.e., of *epistêmê*. [*NE*, Book 6] The second version of the "activity in conformity with a rational principle" that counts as human flourishing is what Aristotle calls politics. Politics is the enactment, in the context of the *polis*, of the rational power of *phronesis*, combined with the reliable display of good character. [See *NE*, Book 10.] *Phronesis* (often translated as 'practical wisdom') is the intellectual power of being able to discern correctly what to do in a particular situation. [*NE* Book 6] It is a cognitive excellence. In contrast to the rational powers of *nous* and *epistêmê* of which *sophia* is composed, however, *phronesis* by definition pertains to phenomena that do ***not*** have form – circumstances, usually. Good character, meanwhile – or the having of a good *hexis* – is a matter of consistently having the right affective reactions to unfolding events, including having an overarching desire to behave virtuously in any given situation. [*NE*, Book 2, especially.] It is a kind of emotional excellence, rather than an intellectual one, although it presupposes that one has accurately assessed the nature of the situation to which one is reacting.

Aristotle calls the combination of *phronesis* and good character (plus a bit of what he calls "cleverness," to help pull off the wise course of action) 'full moral virtue.' The *phronimos* has the right affective reactions, including the generalized desire to pursue the best outcome, but also has the rational power to determine what that course of action is, should it not be readily apparent. Full

moral virtue is contrasted with 'moral virtue,' sometimes rendered 'natural moral virtue.' [*Nicomachean Ethics*, 1144b] (Mere) moral virtue is exhausted by the having of a good *hexis*. One does not have to be wise to have reliably appropriate affective reactions to the circumstances within which one finds oneself, and to wish to pursue correct courses of action in given conditions. Indeed, one need not necessarily be wise even to ***succeed*** in acting well in a given context: many situations are such that if one's habits are sound, one's behavior can be virtuous without one having to make any deliberate choices about how best to proceed. Moreover, even in a situation in which wisdom ***is*** called for, one might get lucky and pursue the correct course of action simply as a reflexive matter of good character. In such a case one might conceive of the display of (mere) moral virtue as being a loose analogue to a Gettier case, in that a habitual response could potentially replicate superficially or accidentally what, for the *phronimos*, would be (when needed) a deliberate deliverance of *phronesis*. As Aristotle puts it, with perhaps a less positive spin, "people may perform just acts without actually being just men, as in the case of people who do what has been laid down by the laws but do so either involuntarily or through ignorance or for an ulterior motive." [NE 1144a]

We are going to assume, for the purposes of argument, that an AI could behave in a way that would be consistent with the having of (mere) moral virtue. A morally virtuous AI would not have affect, which is integral to a good *hexis* in humans, but in lieu of habitually established correct reactions tied to appropriate ***emotions*** an AI (presumably) could be designed to act in a way that, functionally, would be in accord with the mean that Aristotle invokes in Book 2 of the *Ethics*. Such a system might, for example, be trained on a corpus consisting of a large set of morally praiseworthy decisions in such a way that it consistently met expectations with respect to novel circumstances. Obviously, such a system would represent an extraordinary technical achievement and, again, we are not committing ourselves to the view that this is possible; we are only granting it for the sake of

argument. Our thinking in allowing this much is that habituated responses are regular, and as such may admit of being reduced to a machine learning model of the usual sort. After training, the system would act in accordance with the general principle that "given x circumstances, the proper response is to do y (with some probability *n*)."

But (mere) moral virtue only gets us so far. Again, Aristotle is explicit that "natural virtue" is not "virtue in the full sense" *NE 1144b* (i.e., natural virtue plus *phronesis*), and one reason why he thinks that the former is not only ***different*** from the latter, but lacking, is that practical circumstances, according to Aristotle, are particular and always-changing; they are context specific, not universal and not derivable from a fixed training set of instances. As such, consistently excellent moral action requires the capacity to know when a rule should be broken, and what should be done instead. Unlike the unwise but morally good agent, the *phronimos* is capable of determining what to do in ***any*** situation, not just routine ones in which a habituated response, or even a general principle, will suffice. Why? Because in choosing a course of action, the *phronimos* relies upon a well-developed capacity for wise, particularistic judgement, and not – or at least not just – upon the moral equivalent of muscle memory.

Rosalind Hursthouse observed in her original *SEP* entry on virtue ethics that we can understand the difference between full moral virtue and (mere) moral virtue by "thinking about what the virtuous morally mature adult has that nice children, including nice adolescents, lack." (2003). All (by stipulation) want to do the morally correct thing. However, the nice children are not (yet) able to correctly discern, in each unique circumstance, what that might be. (Aristotle claims further that it is the having of *phronesis* that unifies the specific virtues, ensuring that one has all of them.) Our view, for reasons that we set out below, is that an AI will not be able to do better than Hurthouse's nice adolescents. We will call it *nice teenager level* morality (NTL), although we do not mean by the term anything different from moral virtue or natural moral virtue. Achieving NTL

would, of course, be a remarkable technological achievement and, being a tad optimistic about the prospects of this kind of technological development, we can imagine that this may be precisely where machine learning-driven AI is headed as far as its moral capacities are concerned. This said, if one is worried about the prospect of a superintelligence, the idea that is restricted to NTL moral judgment should not be a consolation.

## 2. Can Machine Learning Systems Be Wise, as per Aristotle?

Let us first address what it would take for an AI to have the moral capacities of a nice teenager. The main obstacle to achieving NTL in machine learning systems will be the development of training sets that are sufficient to the task. Some possible components of a training data for such a system could include a suitably large set of examples of moral decision problems and the corresponding actions that are considered morally virtuous, as well as a diverse range of variations on those examples. It would also be important for the training set to be designed in a way that allows the machine learning system to generalize in ways that comport with the underlying principles of morality, rather than just overfitting to specific examples. Ultimately, a successful NTL would be judged a success insofar as it met expectations reliably. The success of a machine learning system designed to embody moral virtues would depend upon its ability to act in a way that is consistent with commonsense human values.

Another strategy might involve using an existing AI, for instance a large language model (LLM), and training it via a set of commonly agreed upon heuristics. This is the strategy used, for example, in DeepMind's SparrowAI (Glaese et.al 2022). SparrowAI is a chatbot built on the Chinchilla large language models that is trained to follow a set of 19 rules or heuristics (for a discussion of Chinchilla, see Hoffman et al 2022). According to the reports of the DeepMind team it does so relatively reliably, but does not have the capacity to adjudicate between these heuristics in

case they come into conflict. The heuristics themselves were the basis of an additional layer of training and do not become topics that the system deliberates upon. Both SparrowAI and ChatGPT add moral (or legal) heuristics as training data after the LLM is in place as a way to censor the outputs of the model in accordance with commonsense heuristics. Thus, SparrowAI is trained to avoid lying; will not pretend to have a body or a history; will not attempt to build relationships with users; will avoid conspiracy theories and stereotyping; etc. Judging what does and does not count as a conspiracy theory is quite subtle, and beyond the capacity of a contemporary LLM. However, in order to avoid conspiratorial thinking, the designers simply train the LLM on a large set of examples of what are commonly taken to be conspiracy theories. Sparrow AI will censor candidate outputs that look like members of that set of examples. As such, the system does not have a definition of what it is to be a conspiracy theory. Rather, it simply avoids outputs that would probably be considered conspiratorial given what people have previously judged as being conspiracies in the training set.

So far, so good – at least for the sake of argument. But what about ***full*** moral virtue? Could an AI be morally akin to a wise adult, rather than a nice adolescent? We think not. Even if it is trained to extrapolate or generalize from its training set, the AI will only ever be capable of actions derived from a probability distribution over elements of that set or via heuristics or rules that serve as the basis for training, as in SparrowAI, or perhaps via a system of hardcoded rules, as in GOFAI. The fundamental problem is that, if Aristotle is right, there is no rule, or set of conditional if-then sequences that could be coded, for telling what the wise thing to do would be in every possible situation. And again, the problem is not that the rule would have to be too long, and/or too complex. On the contrary, the rule is short and sweet: do what the *phronimos* would do. NTL is simply insufficient for the kinds complex and high-risk decisions that autonomous systems would face, e.g., in combat or in medical contexts. The type of decision-making in question calls for

*phronesis*, for *full* moral virtue. And it is not only that the circumstances of practical life require reasoning that is context-specific; equally significant is that the requisite reasoning may well involve multiple or even competing moral principles, requiring the ability to weight different factors and trade-offs. This type of deliberation is difficult for even human beings to consistently do correctly. We would hesitate to allow a human being who had only NTL moral capacity to determine courses of action in such cases, and the situation is no different in the case of an AI.

Still, we can at least ask what it would it mean for a machine to be wise. Robots and AI assistants are increasingly being used in settings in which they *do* make morally significant decisions, and philosophers have recognized that there is a growing need for machines in such roles to have the capacity to make the right ones. The term 'artificial *phronesis*' (Sullins 2021) refers to methods and techniques that are meant to program a capacity for practical wisdom into a machine. 'Artificial *phronesis*,' as the name suggests, would have to consist of more than the ability to perform specific tasks or make decisions based on a data set. An AI with (artificial) practical wisdom would have to be able to understand and navigate complex situations so as to be able to respond correctly to choices between two or more courses of action, each of which (a) has a moral dimension and (b) potentially conflicts with one or more competing moral principle. For example, an AI might be faced with the dilemma of whether to tell a lie in order to protect the feelings of another, or to tell the truth even though it may hurt another's feelings. In such a scenario, the AI would have to be able to resolve the conflict between honesty and kindness, correctly determining which takes precedence in the particular instance.

Would 'artificial *phronesis*' be adequate to the task? We are doubtful. Deliberation of this kind is difficult even for human beings, which is why Aristotle thinks that young people are not yet capable of it – or even busy or distracted adults, for that matter. In the kindness versus truth dilemma, the *phronimos* must consider not only the immediate effects on the individuals involved, but

also the broader implications for trust and relationships within a larger social group and across time. What makes it possible for the wise human to hit the mark, as Aristotle puts it, is that the wise human has, by stipulation, the rational power of *phronesis* (the exercise of which has been cultivated over time). '*Artificial phronesis*' faces the same problem that it is meant to solve, viz., there is no string of code that works the way *phronesis* does. The artificially wise machine would have to be able to judge correctly not just what to do, but which value(s) take(s) precedence over the others. Machine learning systems, however – unlike wise humans – are not able to assign non-arbitrary relative weights to the competing heuristics or principles that serve as the basis for their training (let alone to do so correctly). As long as this is so, 'artificial *phronesis*' will not be the capacity of real *phronesis* – the having of which enables the *phronimos* to go beyond the fixed parameters of habituated good character.

The following are additional cases intended to illustrate the kind of weighting of values, moral demands and/or normative criteria that, in our view, would be beyond the reach of a machine endowed with 'artificial *phronesis*.' (Indeed, rather than saying that the demands of wise deliberation exceed the capacity of artificial *phronesis*, it may be more accurate to say that it is in the nature of the case there is no such thing as programable practical wisdom as per Aristotle.) Again, developing a system that could go beyond NTL to something approaching full moral virtue would require, at a minimum, that it be capable of judging between competing goals in a nonarbitrary way. Our point is that the ability to rank or classify goals is likely to rely upon rational powers that are not reducible to standard training methods for machine learning. For example, one could imagine a system trained on examples that are intended to conform to a principle of respect for others, while also being trained on examples that are based on the principle of fairness or justice. A being who possesses full moral virtue will be able to rank these principles in terms of their relative importance or applicability in different situations. It is difficult to imagine how a machine learning system could be trained to

perform such a ranking, since it would involve weighing multiple kinds of optimization tasks without there being some master optimization task that would automatically subordinate the other two. Consider the allotment of resources in medicine. In some circumstances one may appropriately place a greater emphasis on respect for the person, while other decisions may require us to place a greater emphasis on fairness or justice across populations. One could imagine a machine learning system being trained to produce outputs conforming to both principles, but how would it determine when to apply one rather than another? Would it have some third training set that governs the decision as to which principle to apply in which case? If so, would that trained capacity also be potentially subject to a ranking?

Or take the case of Gaugin, who judged the pursuit of aesthetic worth to be more important than loyalty to his family. As a result, he abandoned his wife and children in order to travel to Tahiti to paint. How should we judge his ranking of what one might construe as aesthetic value over moral value? As it turns out, some of us might accept his decision given the high value that we place on his aesthetic achievement. But what if he had been a mediocre or even an incompetent artist? Even those who might have excused his failure with respect to his wife and children in the actual case would probably find it blameworthy in the case of an artistically deficient Gaugin. The *phronimos* is able to judge rankings of kinds of value such as Gaugin's in a way that the NTL agent cannot. If we were to apply a straightforward moral calculus to his action then he would clearly have been blameworthy. But there are times when aesthetic value trumps a straight-forwardly moral consideration. And we do not even need to decide the Gaugin case to appreciate the claim. Certainly, if one has to tell a minor lie in order to produce an artistic masterpiece, one is likely to be forgiven their sin by most mature adults (Author).

A third example is a potential conflict between what might be thought of as a moral versus a prudential value. Say I have had a lifelong wish to hear the rock and roll band Primus live in

concert. However, I have promised my colleague that I will water his geraniums every day while he is traveling to a conference. In order for me to see Primus, I would have to travel some distance and stay overnight, thereby violating my promise to water my colleague's flowers. In this case, at least some of us would say that breaking one's moral obligation is forgivable in light of the important prudential value of seeing Primus. How might a machine make such a judgment? To reference a familiar image, we can imagine an adolescent banging their fist on the table and insisting that a promise is a promise and that it cannot be broken. While it may not have a fist with which to bang, we expect that a machine learning system with NTL capacity will be similarly unable to rank the relative importance of the moral and prudential factors in such a case. Ultimately, the ranking of distinct kinds of normative reasons, as discussed for example in (Author) is a capacity that exceeds any set of training instructions.

Machines with even NTL moral capacities would be a remarkable technical achievement. They would, however, suffer from the same kinds of ethical shortcomings that we regularly see in human adolescents when they attempt to deliberate. One weakness of adolescent moral reasoning is that it is vulnerable to the influence of peers. Adolescents are also prone to simplistic or black-and-white thinking, as well as to a tendency to focus on the immediate consequences of an action rather than its broader implications. An AI possessed of a necessarily artificial version of (mere) natural moral virtue would be susceptible to these same deficiencies. By design, machine learning systems aim to conform their outputs to the normal or probable characteristics of their training data. And, as we have already noted, a machine learning system has no capacity to reflect critically upon the principles governing its training or on the relative merits of its training data. We have already discussed inability of such a system to correctly assess irreducibly particularistic circumstances so as to reliably determine the correct course of action in any given situation. But we can see also that

such a system would, in virtue of its design and the limitations thereof, lack perspective and be guided almost entirely by a spirit of conformism.

## 3. A Word On the Meta-Philosophical Implications of Full Moral Virtue

If machines in principle cannot have Aristotelian full moral virtue, that fact will be important for those who are thinking about ethics and AI. Of course, it will be significantly less important if an approach to morality that might be more easily codified turns out to be the correct one. Little of what we have argued will be conclusive for the Kantian or the utilitarian, say, neither of whom would recognize the theoretical need for the distinction between NTL and the *phronimos*. As it happens, we doubt that either Kantianism or utilitarianism ***are*** superior to Aristotelianism, so for us it matters, when we reflect upon machine learning, that *phronesis* does not look to be programmable or trainable via conventional machine learning techniques. But we also take our conclusion to be illustrative of a meta-philosophical point regarding metaphysical consistency, both between moral philosophy and metaphysics, and within metaphysical theories themselves. Paradoxically, the technical façade of AI ethics encourages a kind of mix-and-match permissiveness in this regard, in our view. There is also an unfortunate lack of attentiveness to implicit categories or assumptions, if only these are deeply enough embedded in a given account. The metaphysical distinction between *phronesis* and habit – or, if you prefer, between habit plus *phronesis* (full moral virtue) and habit alone (good character or moral virtue) – places a limit not just upon what kind of entity can be wise as per Aristotle, but on what may and may not be combined metaphysically with a belief in the rational power of *phronesis*.

We have dubbed the dominant, contemporary analytic metaphysics 'nomological.' It is so named because of the decisive role that is played in it by regular sequences (be they conceived as deterministic or as probabilistic) when it comes to the ontology of causation, and by extension the

ontology of agency (see, e.g., Ellis 2001, Mumford 2004). The label 'nomological metaphysics' includes contemporary Humeanism, but also those contemporary metaphysical accounts that derive from Kant (and others). The framework in question has various recognizable features, but it is the replacement of Aristotelian powers with a rubric of law (or at a minimum regularity) that is salient for present purposes. However, the nomological approach is no longer an unassailable orthodoxy. Unlike Hume himself, even Humeans – or what (Author), following Brian Ellis, calls 'passivists' – are now coming to affirm the existence of what they call powers (although said powers are conceptualized by passivists in Humean terms, in terms of regular sequences) [Author]. To complicate matters, not only is the nomological approach often equated with or assumed to be required by natural science, it is – as noted earlier – often held tacitly.

Aristotelianism, or perhaps neo-Aristotelianism – is an alternative that has been gaining increasing traction in recent decades. With respect to the metaphysics of causation – including but not limited to the actions of agents – the Aristotelian ontology rests not on laws, but on the expression of powers, viz., capacities for doing, or activity, of one kind or another. [See, e.g., Lamprecht 1967; Anscombe 1993; Harré and Madden, 1975; Bhaskar 1975; Mumford and Anjum, 2011; Author]. Substances and artifacts alike, from this perspective, are thought to have such capacities. Since Aristotelians do not deny the reality of activity, powers need not be thought to reduce to sequences of static states of affairs. So-called causal laws, finally, can be seen to be descriptions of the behavior of things, given their powers, as they interact with other things.

Our own neo-Aristotelian view is that different 'powerful particulars,' to use Harré and Madden's locution, have different kinds of powers, in virtue of which they are able to behave in different kinds of ways. Some powerful particulars at least sometimes behave in regular ways, such that nomological talk about the display of ***their*** powers (sometimes) may be descriptively adequate (albeit incorrect metaphysically), sustaining either deterministic or probabilistic law-statements. But

not all do. (See Author, for a lengthy discussion of powers and regularity). Aristotelian *phronesis*, as we have argued, is a power of moral agents, the expression of which, by definition, ***cannot*** be described in terms of regular sequences – which fact, as we have emphasized, differentiates it from the habitual responses associated with (mere) moral virtue. Of course, we ourselves understand even habitual responses in terms of an Aristotelian metaphysics: habits, like, rational powers such as *phronesis*, are – as we see it – expressions of powers. But habits, unlike *phronesis*, are arguably the type of phenomenon that a nomological metaphysics can accommodate (if only at first blush). The Humean, for instance, may plausibly construe a habit as a specifiable behavioral regularity (they will have trouble saying anything more than that, in our view, but that is neither here nor there), but *phronesis* as a behavioral regularity amounts only to "reliably identifies the correct course of action."

Thus far, our Aristotelian-inflected claim has been simply that the distinction between the power of *phronesis* and the power(s) involved in a habitual response precludes machines from having full moral virtue. We are now in a position to add to this point two meta-philosophical observations. First, for precisely the same reasons that a machine cannot be wise as per Aristotle, any moral theory that takes *phronesis* to be a real phenomenon will be incommensurable with a nomological metaphysics. Why? Because *phronesis* is not a matter of regular sequences, and regular sequences are the building blocks of a nomological metaphysics. (Nor will it help to switch to talk of dispositions, as any talk of *phronesis* being a disposition will either turn it back into a habit or will have to presuppose that dispositionality is something other than a stimulus-response relation, once again exceeding a nomological framework.) None of this poses any difficulty for Aristotle himself. Indeed, he is the author of the very metaphysics that is required by his moral theory. But the same cannot be assumed of contemporary philosophers.

Second, in saying that a commitment to the reality of *phronesis* requires that one abandon a nomological ontology – or any ontology that emphasizes metaphysical completeness, for that matter

– we are also saying something to those who want to retrofit powers into a nomological metaphysics. At a minimum, that is, we are saying that a metaphysics that admits of the reality of the power of *phronesis* cannot be one in which powers are equated, either directly or via the phenomenon of a disposition (if one thinks that powers and dispositions are different), with sequences in any guise. We say "in any guise" because sequences – be they necessary or contingent, deterministic or probabilistic – seem always to be lurking at the back door of contemporary metaphysics, if not at the front. At this level of abstraction, then, what we learn from the fact that machines cannot be wise as per Aristotle is that a metaphysics of powers that allows for *phronesis* cannot be made to sit atop of nomological categories any more than can an Aristotelian ethics.

**Bibliography**


Anscombe, G. E. M. (1993) "Causality and Determination." In E. Sosa M. Tooley (ed.), *Causation.* Oxford: Oxford University Press.

Aristotle. (1962) *Nicomachean Ethics*, Translated, with Introduction and notes, by Martin Ostwald. Indianapolis: Bobbs-Merrill Educational Publishing, 1980 printing.

Bhaskar, Roy. (1975) *A Realist Theory of Science*. Leeds: Leeds Books.

Ellis, Brian. (2001) *Scientific Essentialism.* Cambridge: Cambridge University Press.

Glaese, A., McAleese, N., Trębacz, M., Aslanides, J., Firoiu, V., Ewalds, T., & Irving, G. (2022). "Improving alignment of dialogue agents via targeted human judgements." *arXiv preprint arXiv:2209.14375*.

Groff, Ruth. (2013) *Ontology Revisited: Metaphysics in Social and Political Philosophy*. Oxon and New York: Routledge.

Groff, Ruth. (2019) "Sublating the Free Will Problematic: Powers, Agency and Causal Determination." *Synthese*, Volume 196: 179-200.

Groff, Ruth. (2021) "Conceptualizing Causal Powers: Activity, Capacity, Essence, Necessity. *Synthese*, 199 (3-4): 9881-9896 Topical Collection, "New Foundations of Dispositionalism," edited by Andrea Raimondi and Lorenzo Azzano. https://doi.org/10.1007/s11229-021-03229-x

Harré, Rom and Edward H. Madden. (1975) *Causal Powers: A Theory of Natural Necessity*. USA: Rowman and Littlefield.

Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., & Sifre, L. (2022). "Training Compute-Optimal Large Language Models." *arXiv preprint arXiv:2203.15556*..

Hursthouse, R. and Glen P. (2022) "Virtue Ethics", *The Stanford Encyclopedia of Philosophy* (Winter 2022 Edition), Edward N. Zalta & Uri Nodelman (eds.), forthcoming URL = <https://plato.stanford.edu/archives/win2022/entries/ethics-virtue/>.

Lamprecht, Sterling P. (1967) *The Metaphysics of Naturalism*. New York: Appleton-Century-Crofts.

Mumford, Stephen. (2004) *Laws in Nature*. New York and Oxfordshire: Routledge.
Mumford, Stephen and Rani Lil Anjum. (2011) *Getting Causes From Powers*. Oxford: Oxford University Press.

Sullins, J. P. (2021). "Artificial Phronesis." *Science, Technology, and Virtues: Contemporary Perspectives*, 136.

Symons, J. (2015). "Physicalism, scientific respectability, and strongly emergent properties." In Dima, T., & Luca, M. (Eds.). *Cognitive Sciences: An Interdisciplinary Approach*. Pro Universitaria., 14-37.

Symons, J. (2018). "Brute facts about emergence." *In* Vintiadis, E., & Mekios, C. (Eds.). *Brute facts*. Oxford University Press.

Symons, J. (2021). "Meaningfulness and kinds of normative reasons." *Philosophia*, *49*(1), 459-471.