Virtue and artificial moral intelligence

Marinus Ferreira[1]

Abstract. As AI decision-making plays an increasing role in our daily lives, and makes more and more important contributions to how we manage our most vital interests, the question of to what extent algorithms can be made sensitive to the full scope of our moral concerns is of ever greater concern. My proposal is that we can identify a scale along which we can place the suitability of an AI system for *virtue judgements*, meaning appropriately evaluating instances of virtue (or vice). To do so, I harness the notion of a developmental pathway, where virtue theorists describe the acquisition of virtue as a task involving movement along a number of stages of increasing sensitivity to and autonomous control over the facets that go into virtuous action. This prominently includes not just behavioural capacities (i.e. a carpenter building a table, an AI succeeding at a sorting task) but also psychological capacities, such as perception, imagination, foresight, and so on. We need to have multiple levels of evaluation in order to make sense of virtue judgements: that of how the action in question fares in this particular case, and how the disposition that this action is an instance of fares in the larger context of the agent's life. We need appropriate interpretation of the subject matter of our judgements (not just as arrays of formal symbols, but mapped onto tangible situations) in order to appropriately judge the relationship between an individual action and the disposition which it is an instance of, since we need a sensible understanding of how having concrete disposition of character is of substantial import on concrete actions of the respective kind, something that involves more than formal relationships but rather an appreciation of the referents of the numbers and relations that go into machine learning.

1 Introducing the analogy between human and machine learning of virtue

Given the prominent importance of virtue and vice in human moral frameworks (in that every human society has such a framework [8]) determining to what extent AI can be made sensitive to virtue and vice goes a long way to determining to what extent they can be sensitive to the full range of human moral sensitivities. The most promising avenue for pursuing the prospect that artificial intelligences could develop genuine virtue is to draw a link between the way that a machine learning agent becomes trained and the way that human agents are meant to develop virtue. In both cases, there is the repeated engagement with the target domain, and the incremental improvement of the agent's sensitivity to the relevant features of the domain is what drives learning, leading to responses which are better fitted to the domain, and more reliably reaches the target states. The more we can cast the training of a machine learning agent in the lights of the development of virtuous character traits, the more compelling this analogy becomes.

[1] Centre for Ethics in Public Life, University College Dublin, email: marinus.ferreira@ucd.ie

An obstacle to the analogy is that the usual way we understand a human agent developing virtue is tied to the inherent qualities of the features that are being trained, i.e. courage has to do with the regulation of fear, an inherent psychological feature of humans. The more we flesh out what this development is meant to consist in, the more we talk about the specifics of human psychology. For instance, Aristotle in his extended discussion of courage [4] compares it with a different, related psychological feature, the disposition of confidence. The difference between courage (which is a full-blown virtue) and confidence (which is a valuable disposition for people like soldiers to develop, but falls short of a virtue) depends on such things such as that it is a psychological truth about people that they are more likely to remain steadfast in what they are doing the more the situation they are actually in corresponds to the situations they have trained to be in. Neural networks and other machine learning agents don't have psychologies, or at least not of the kind that these kinds of psychological truths latch on to, and as such the kind of fine-grained discussion Aristotle and those following him engage in doesn't have an equivalent.

2 One-dimensional vs multi-dimensional judgements

With the above comparison between human and machine learning in mind, the question becomes whether there is enough in common between the training of human and machine learning agents in order to maintain the analogy between the two, such that we can sensibly talk about machine learning agents as learning virtue. I believe we can definitively answer this question in the negative, but not with the kind of consideration that we've surveyed thus far. We need to change the focus from how machine learning agents may or may not produce behaviours in a relevantly similar way as human agents, to investigating the way that the agent-based antecedents of virtue leads to virtuous behaviours. The clearest example of such an antecedent would be causal antecedents, and the claim that the relationship between an agent's mental states and their action is causal is something like the standard view in the philosophy of action, and I myself elsewhere endorse a functionalist version of this pertaining to the production of virtue [7]. I want to split the way we describe action into two descriptive profiles: a *behavioural profile* which describes the movements and other first-order physical manifestations of the action which are contained situationally in the action, and a *agential profile* which describes the so-called 'inner' features of the agent involved in the production of the action but which are present cross-situationally, which in the human case are mental states, intentions, perceptual sensibilities, and various other relevant dispositional qualities. The domains where there is the greatest interest in artificial intelligence there is no gap between the behavioural profiles of competent human and competent machine learning agents, such as where we are interested in

the advice that we would receive for particular problem cases (like diagnostic systems) or the action in question is one that computers can perform at least as well as humans (like electronic trading systems). The question is then how we can compare the agential profiles of human and machine learning agents when it comes to their producing the behavioural profiles in question.

The feature of agential profiles I want to latch on to is the richness of their relationship to behavioural profiles. In particular, I want to distinguish *one-dimensional* to *multi-dimensional* relationships. A one-dimensional approach would be one which can be exhaustively described with a mapping from a domain of situations to a range of behaviours. I describe this as one-dimensional because there is just one way that a target situation and the respective behaviour is related, and that is by something like a 'to-be-done' relation from one to the other. A multi-dimensional approach, then, is one which cannot be exhaustively described by such a relationship, but instead will have some cases where the same situation described in the relevant amount of detail does not prompt a specific behaviour, but instead would prompt different behaviours based on different features not specific to the situation, but (as it is in the majority of examples) due to (cross-situational) features of the agent.

The claim isn't that a one-dimensional approach involves only a single parameter whether the behaviour in question is 'to-be-done' for the specific situation, since such an approach would be crazy even without considering the realities of machine learning and its thousands of parameters (OpenAI's GPT-2 language-parser, aimed at understanding English, has over 1.5 billion parameters, as an extreme example [15]). The question is instead about how the output of the decision-making process (human or otherwise) relates to the target behaviours. A number of theories of great sophistication, rational choice theory and consequentialist ethics most prominently, involves one-dimensional judgements as I have described them: for instance, even the most sophisticated versions of these two (mutually reinforcing) decision-making approaches hinges on an operative move where the available options are ranked according to expected utility (or some similar metric), and the top-ranking option is the one that is 'to-be-done'. All the sophistication of the theories is contained in the agential profile that they recommend for you; the appeal of these theories is that this sophistication need not get in the way of the simple and powerful relationship between situations and recommendations these theories offer (see, for instance, [13] [14]). We can use my approach as a way to articulate one important strain of criticism against these theories, that they require an overly impoverished view of what decision-making in these domains involve, and that the value of these kinds of (what I call) one-dimensional judgements is illusory [21, 17]. Whatever the standing of those objections to consequentialism and rational choice theory, my claim here is that virtue judgements cannot appropriately be described as one-dimensional.

2.1 Examples of where one-dimensional judgements are insufficient

We can give many examples of the one-dimensional/multi-dimensional split from the literature, because the kind of relationship it describes in ways to relate behavioural and agential profiles is of general interest. There are many examples of discussions of virtue judgements which falls on the wrong side of the split, such as are various objections to virtue ethics or versions thereof which fail because they mistakenly describe virtue judgements as one-dimensional. The most widely known example is the view that was commonplace in the middle of the 20th century (and which we find in, for instance, Rawls [16]) which treats virtues as if they were nothing more than the disposition to behave in virtuous ways. This approach has now fallen decisively out of favour, because the lynchpin of the development of contemporary virtue theories has been a renewed focus on how the way that virtuous agents engage with the world is qualitatively different from how agents lacking in virtue do. One important example is John McDowell's claim that virtue changes the perceptual features of an agent, such that an agent with a virtue sees thins that gives them reasons that someone without virtue do not see nor have [12]. A simple example we could use here is of someone like a Boy Scout who wants to do their good deed for the day, and wanders around looking for one to do, walking straight past the elderly person hesitating before crossing the street because they don't trust their ability to cross quickly enough to avoid being hit by a car, straight past the young parent whose bag of groceries has burst and is trying to simultaneously gather their goods while minding the young child in their care, etc. The bumbling Boy Scout fails to respond to these cases because, lacking the virtue of kindness, they do not recognise these as opportunities to be kind and that these are occasions to do their good deed for the day. This is meant to just be a failure of perception, rather than a lack of motivation or false views about what kindness and good deeds consist in. This case is multi-dimensional rather than one-dimensional because each virtue in question has its own attending perceptual capacities, and its own domain of reasons that prompts virtuous behaviours which fail to trigger the appropriate behaviour in agents lacking virtue. It isn't that the bumbling Boy Scout is failing to recognise the to-be-doneness of helping the elderly person across the street or helping the young parent collect their goods, because plausibly these actions are not brutely to-be-done. It is that the Boy Scout fails to recognise these are instances of kindness, and only indirectly that these are potentially things that are to-be-done.

Another example from the virtue literature of where multi-dimensional judgements are required is the response among virtue theorists to the so-called situationist challenge, the view that there couldn't be virtues because results from social psychology speak against the necessary cross-situational agential features (that make up the agential profile in my account) are not robust enough to feature as plausible explanations for behaviour. The way the challenge is meant to work is that there are many instances where an agent is faced with a choice between what appears to be a virtuous action and a non-virtuous one (e.g. helping a conspecific pick up a bunch of papers that have been dropped, or rushing past without helping), very minor changes to the situation can lead to radical differences in the rate to which participants to the study perform the virtuous action (in this case, whether they found some spare change in the coin return slot of the phone box they've just used). Because the minor situational features seem to swamp out other influences in the action, the challenge goes that virtues (or any other personal, cross-situational disposition) could explain the behaviour. This challenge has been roundly rejected by the vast majority of virtue theorists, and on a number of different grounds, most often that the experiments fail to test for virtue. I want to highlight Nancy Snow's response to the challenge, as one that engages with the relevant psychological literature to a great extent [18]. In addition to various criticisms of the experiments and their interpretation, Snow stresses that to conclude that these experiments undermine our belief in cross-situational character traits such as virtues and vices requires us to have too shallow and restricted an understanding of how individuals engage with the situations they find themselves in. The situationists' interpretations are what I call one-dimensional, in that they code an individual's

response to a situation by mapping from a situation to a bare behavioural (in our example, either as according to virtue when they help pick up the papers, and not according to virtue when they don't). Following an approach prevalent in the psychological literature, she phrases this in terms of objective construals (how actions are understood when viewed as much as neutral behaviours as possible) as opposed to subjective construals (how agents understand the situation and the actions themselves). Snow points out that many studies, prominently including follow-up studies to famous experiments harnessed by situationists such as the Milgram experiment, agents most frequently have multi-dimensional (in my terms) subjective construals of the situations such that they maintain cross-situational consistency in the way that the one-dimensional (in my terms) interpretations of the situationists obscure. The situationist interpretation, by being one-dimensional, is rejected as not capturing the operative features of the decision-making at issue in these debates (In my terms the objective construal is the same as the behavioural profile, but the subjective construal is only a part of the intentional profile; for a longer discussion, see [7].)

The relationship between specific action-types and some overall to-be-doneness is Elizabeth Anderson's pluralism which is not a virtue-based approach, but can be made amenable to one [1], and articulates features of a pluralistic virtue theory like Christine Swanton's [19]. In her criticism of monistic theories (especially rational choice theory and consequentialism), Anderson highlights that even in the cases where we can see some specific values as components of some overarching, all-encompassing value (which may be monistic), it is irrational and self-defeating to pursue the overarching value without regard to the plural component values. This is because the overarching value is a product of the component values, and if the component values were to wither away, the overall value would as well. It is one thing to believe that it is rational to make trade-offs between component values to best promote the overall value, but it is another to treat the component values as disposable, because if they go away the overall value goes as well. As such, Anderson warns against this mistaken inference from there being a measure of overarching value to there being nothing to measure except the extent of overarching value. This means that even judgements of overarching value, if such a monistic measure is available, is an example of multi-dimensional judgement, because what the judgement of the overarching value consists in is the combination of multiple dimensions of judgement, corresponding to each component value in play.

3 Why machine learning is not like learning virtue

Having introduced one-dimensional vs multi-dimensional judgements and their import, to virtue and to ethics in general, let us move on to describing why there is a mismatch between machine learning and what is required to learn virtue.

3.1 Learning virtue is multi-dimensional

When I describe how learning virtue is meant to work, my main source is Julia Annas in her contributions to the mainstream, neo-Aristotelian view of the virtues and virtue ethics, especially her paper 'Being Virtuous and Doing the Right Thing' [2]. Her task there is to explain how virtue judgements are meant to be different from what she calls the 'technical manual model' of ethics, where the final end of moral theory is the production of a set of guidance, like a decision procedure, that requires no special ethical expertise to follow, and results in the performance of morally right behaviour. For our purposes, it is important to note that the technical manual model would result in only one-dimensional judgements in my terminology, since this manual would consist exactly in a mapping between situations and behaviours, and the lack of a requirement to have ethical expertise to follow such a manual just is the requirement that cross-situational agential features are not in play. In contrast to this, Annas offers a developmental approach, where there are at least two tasks of moral philosophy: telling us how to develop as ethical agents (the 'being virtuous' part of her title), and what behaviour is distinctive of well-developed ethical agents (the 'doing the right thing' part).

The main thing Annas focuses on is how the two different tasks in virtue judgements involve two different kinds of evaluation: evaluating the action as it stands on its own (which the technical manual model can capture), and evaluating the action as part of a pattern of actions over a period of time (which it cannot). She notes the most prominent way to try and introduce these factors into contemporary theory discussion, Hursthouse's qualified agent model: right action is doing what the virtuous agent would do [11]. She straight away stresses that Hursthouse did not mean this to be a decision procedure, and that she and Hursthouse agrees that it could not be one. Some treatments of virtue ethics wrongly take 'what the virtuous agent would do' as what is meant to define right action [20]. But Hursthouse does not believe this; Hursthouse thinks right action is action that constitutes *eudaimonia* (flourishing as a human being; she doesn't consider machine agents, and non-human animals can flourish but don't have the rational capacity for full *eudaimonia*). Hursthouse's proposal is an answer to 'what do I do in these situations', not to 'how do I develop as an agent'. She uses her qualified agent account to show that our knowledge of the virtues on its own informs right action, and does not need to be a criterion of right action to do so. In short, virtue judgements are multi-dimensional in the Hursthouse and Annas accounts, and not reducible to one dimension.

The challenge then is to try and find an informative way to take us from unqualified reasoning to qualified reasoning. Annas does so by introducing the notion that our reasoning falls within a developmental pathway. This is part of her wider analogy between virtue and skill [3]. At both the beginner and expert ends we're still doing the same thing towards the same ends. But, as we move along the pathway, the manner we do so changes. This is not to claim that the action-guidance given by expert reasoning differ from the action-guidance given by beginner reasoning . Perhaps they do, but the extent it does so is sharply limited (or, to use the terms used above, expert and beginner reasoning won't drastically differ about what is to-be-done). If you ask a beginner carpenter and an expert carpenter to build a table, you would like a table at the end in both cases. A difference in the end product might come about if the expert does something the beginner is unable to do, but not otherwise: both the expert and beginner should build you a basic table. Mainly you'd expect the differences to be in the way they build the table differently from the way the expert does: taking more time and care to get basic things right, for instance, or that they would be less sensitive to features that make a difference to how good a table it would be. Thinking about the beginning carpenter, when we look at how they go about building the table, there are two things we're looking at: firstly, how this action contributes to the end (building the table); and secondly, how this action contributes to development (improving as a carpenter). When someone builds a table, they do both these things, and we care about how well they do both these things. Both these things are objects of evaluation, and they are also different evaluations. Here we have multi-dimensional judgements, which is required to move us along a developmental pathway.

The Annas point is that these two tasks, learning to be virtuous and learning to do the right thing, are different but mutually supporting. Succeeding at one task makes it more likely that you will succeed at the other. You get the right personal traits, the right agential profile, by getting a lot of experience at carpeting, such as building tables, etc. Those traits are the right ones because they are the ones that make you succeed at building tables, etc. So, we have a process of scaffolding between the agential and behavioural profiles. The best way to develop your agential profile is to practice at displaying a particular behavioural profile (for a carpenter, building a lot of tables requiring the upper reaches of their current skill level). As the agential profile develops, so too does the sophistication of the behaviour profile that can be displayed (for a carpenter, what counts as a table that requires the upper reaches of their current skill level). And as the behavioural profiles of the individual actions the agent displays, this in turn leads to further development of the agential profile. We have mutual support between the different aspects of the multi-dimensional judgement.

The qualified agent account, taken on its own, does not provide a decision procedure, largely because it does not give us a way to relate ourselves to the qualified agent. It also only gives you the first kind of evaluation, involving behavioural profiles, and as such is one-dimensional. Adding a developmental pathway gives us the required relationship between behavioural and agential profiles. The decisions of the qualified agent are what someone at the end of such a developmental process would act like. People not yet at the end of the pathway need to consider both what the qualified agent would do, and both how their reaction to a scenario would need to change to be like that. And, of course, almost everybody is only part of the way on the process. As for carpentry, so for virtue, on Annas's skill analogy.

3.2 Machine learning makes for one-dimensional judgements

Machine learning agents have shown great skill at learning to perform a wide variety of complex tasks, normally by repeatedly testing strategies in the target domain against some metric of success. What is important about this process for our purposes is that the relationship between the behavioural and agential profiles is one-dimensional. At the moment the trend in AI development (as in deep learning algorithms) is towards systems that treat the domains they work on as uninterpreted: medical diagnosis algorithms deal with sets of numbers, not features of patients and diseases, and natural language processes increasingly don't treat communication as linguistic items but as statistical relationships across entries in a database. This immediately presses the relationship between the agential and behavioural profiles towards being one-dimensional. In at least the human case, and quite plausibly it must be like this in general, one of the most important ways in which the agential profiles of the agent becomes richer is through an engagement with the specificities of the target domain and the objects within in (see the extended use of this device in, for instance, particularistic theories [6, 10], including the particularistic virtue ethics of Swanton [19], and the imagination-based virtue ethics of Sophie-Grace Chappell [5]).

Deep-learning algorithms and the like collapses its engagement with its subject-matter to a single level, that being the mapping between the inputs the algorithm considers and the metrics by which its outputs are judged, such as a natural language processing algorithm taking a string of text and determining which text token is most probable to occur next, given the (uninterpreted) statistical model it has developed through its training. It can have an extremely sophisticated model, sensitive to conditional features, but in the relevant respect there is still just one level of evaluation: in this example, what token is most likely to occur next. The machine learning agent (or, at least, its developers) are of course immensely interested in how through machine learning can become more sophisticated in its judgement, but this doesn't make the judgement in question multi-dimensional. It isn't that there is a firm developmental pathway which the machine learning agent may deviate from, but this is exactly what happens in the human case (some virtue theorists, like Philippa Foot, take vice simply to be deviation from the proper development as a human [9]). The point of machine learning procedures is that there isn't such a developmental pathway, there is just the repeated measure of attempts by the machine learning agent against the given metric of success. The improved performance according to this unchanging metric is a one-dimensional judgement, not like the mutual scaffolding of two different judgements as we have in the interplay between behavioural and agential profiles.

4 A scale of suitability for learning virtue

I do not rule out that AI systems can develop their own distinctive intentional profiles which fulfils the same role as human psychology but without mimicking the way humans do it. This prospect is as yet far-fetched, and as argued above the reliance on machine learning in its currently popular form stunts the prospects of it developing, but in what follows I describe how we can incorporate this possibility, by sketching out a scale of suitability for virtue judgements. The basic idea behind the scale is that it is a higher-order version of the developmental pathway approach of Annas et al, where me measure progress on this scale not by way of how much development has actually occurred, but instead by the potential an agent has for moving along the stages of the developmental pathway. A challenge for such an approach is not to make it too parochial, by deciding that what makes you have more potential for virtue development is to be more like a human agent. We need to accommodate the fact that all paradigms we have of an agent moving along a developmental pathway are human agents, without making it out as if it could only ever be human agents (or perhaps extraterrestrials who are very much like humans except in their origin, like Star Trek aliens). Since the way I have distinguished between human and machine learning agents by way of the richness of the relationship between the behavioural and agential profiles has been at a high level of generality and has not referred to concrete features of human psychology (of the kind surveyed when we first evaluated the analogy between the training of human and machine learning agents), my hope is that it can be harnessed to produce at least a plausible starting point for the development of such a scale.

As discussed above, machine learning agents have (at least in principle) no difficulty in the performance of some given behavioural profile. It is the ability of machine learning agents to develop extremely sophisticated and successful performance in a wide variety of domains that drives much of the interest in machine learning, after all. What would be needed to move towards multi-dimensional judgements would be something like the component values as we find in my discussion of Anderson above. There would need to be a range of self-sufficient domains of value, all of which the machine learning agent must undertake evaluations of fully in their own right. If there is some overarching value which is a product of all of these components, then that overarching value must not crowd out the judgements of the component values. This would amount to one further step on the scale of suitability for virtue judgements (and, as

Anderson stresses, any non-monistic judgement). At least one further steps would be needed. There would also need to be a developmental pathway that the machine learning agent should be on, such that deviation from that pathway at least makes vicious behaviour more likely (if it doesn't outright constitute vice). There is no prospect of spelling out such a pathway in a venue like this, despite it being of great interest. Here I must content myself with only describing it in the barest outlines. One thing we know it must engage with, if there is to be any comparison between virtue in human and machine learning agents, is that the developmental pathway must at least measure the extent of sensitivity to the kind of qualitatively different reasons that are distinctive to virtuous reasoning, following McDowell. That is, it must not just involve more fine-grained sensitivity to the same features as an agent at the beginning of its training has, but distinctly different features must be represented in its analogue to its perceptual capacities. As long as machine learning uses the same metric of success throughout its training, it will not be possible to attain this third step on my proposed scale of suitability for virtue judgements. How such qualitatively different capacities are to be developed is a question I for one am greatly interested to see answered.

REFERENCES

[1] Elizabeth Anderson, *Value in ethics and economics*, Harvard University Press, 1995.

[2] Julia Annas, 'Being virtuous and doing the right thing', *Proceedings and Addresses of the American Philosophical Association*, **78**(2), 61–75, (2004).

[3] Julia Annas, *Intelligent Virtue*, Oxford University Press, Oxford, 2011.

[4] Sarah Broadie and Christopher Rowe, *Aristotle: Nicomachean Ethics: Translation, Introduction, Commentary*, Oxford University Press, Oxford, 2002.

[5] Timothy Chappell, *Knowing what to do: imagination, virtue, and platonism in ethics*, OUP Oxford, 2014.

[6] Jonathan Dancy, *Ethics without principles*, Oxford University Press, Oxford, 2004.

[7] Marinus Ferreira, *Limited Conventions about Morals*, Thesis, 2017.

[8] Owen Flanagan, *The Geography of Morals: Varieties of moral possibility*, Oxford University Press, Oxford, 2016.

[9] Philippa Foot, *Virtues and Vices*, Oxford University Press, Oxford, 2nd edn., 2002.

[10] John F Horty, *Reasons as defaults*, Oxford University Press, Oxford, 2012.

[11] Rosalind Hursthouse, *On Virtue Ethics*, Oxford University Press, Oxford, 1999.

[12] John McDowell, 'Virtue and reason', *The monist*, **62**(3), 331–350, (1979).

[13] Philip Pettit, *The Consequentialist Perspective*, 92–174, Blackwell, Oxford, 1997.

[14] Philip Pettit, *The robust demands of the good: Ethics with attachment, virtue, and respect*, OUP Oxford, 2015.

[15] Alec Radford, Jeffrey Wu, Dario Amodei, Daniela Amodei, Jack Clark, Miles Brundage, and Ilya Sutskever, 'Better language models and their implications', *OpenAI Blog*, (2019).

[16] John Rawls, *A theory of justice*, Harvard University Press, Cambridge, MA, 2009.

[17] Amartya K Sen, 'Rational fools: A critique of the behavioral foundations of economic theory', *Philosophy and Public Affairs*, 317–344, (1977).

[18] Nancy E Snow, *Virtue as social intelligence: An empirically grounded theory*, Routledge, 2010.

[19] Christine Swanton, *Virtue ethics: A pluralistic view*, Clarendon Press, Oxford, 2003.

[20] Mark Timmons, *Moral theory: an introduction*, Rowman and Littlefield Publishers, 2012.

[21] Bernard Williams, 'A critique of utilitarianism', *Utilitarianism: For and against*, **77**, 108–118, (1973).