AI & SOCIETY
https://doi.org/10.1007/s00146-021-01325-7

ORIGINAL PAPER

Artificial virtuous agents: from theory to machine implementation

Jakob Stenseke[1]

Received: 29 May 2021 / Accepted: 15 November 2021

Abstract
Virtue ethics has many times been suggested as a promising recipe for the construction of artificial moral agents due to its emphasis on moral character and learning. However, given the complex nature of the theory, hardly any work has de facto attempted to implement the core tenets of virtue ethics in moral machines. The main goal of this paper is to demonstrate how virtue ethics can be taken all the way from theory to machine implementation. To achieve this goal, we critically explore the possibilities and challenges for virtue ethics from a computational perspective. Drawing on previous conceptual and technical work, we outline a version of artificial virtue based on moral functionalism, connectionist bottom–up learning, and eudaimonic reward. We then describe how core features of the outlined theory can be interpreted in terms of functionality, which in turn informs the design of components necessary for virtuous cognition. Finally, we present a comprehensive framework for the technical development of artificial virtuous agents and discuss how they can be implemented in moral environments.

Keywords Machine ethics · Virtue ethics · Artificial moral agent · Moral machine · Connectionism · Artificial neural networks · Artificial intelligence

1 Introduction

As artificial systems enter more domains of human life, the last decades have seen an explosion of research dealing with the ethical development and application of AI (AI ethics), and how to build ethical machines (machine ethics).[1] While efforts of the former kind have seemingly converged on a set of principles and guidelines (Floridi and Cowls 2019), their capacity to have any substantial impact on the ethical development of AI has been called into question (Hagendorff 2020; Mittelstadt 2019). Lacking mechanisms to enforce their own normative claims, AI guidelines might instead serve as "ethics-washing" strategies for institutions. To bring ethics into the very core of the research and development of AI systems, it has instead been suggested that ethicists should adopt the role of designers (Van Wynsberghe and Robbins 2014). Accordingly, several attempts have been made to implement ethical theory in machines, with the majority of them taking one of three approaches: consequentialism (Abel et al. 2016), deontology (Anderson and Anderson 2008), or hybrids (Dehghani et al. 2008).

Virtue ethics has several times been proposed as a promising recipe for artificial moral agents (Berberich and Diepold 2018; Coleman 2001; Gamez et al. 2020; Howard and Muntean 2017; Wallach and Allen 2008). Due to its emphasis on moral character and moral development, it offers a path to equip artificial moral agents (AMAs)[2] with the ability to learn from experience, be context-sensitive, adapt, and conform to complex human norms. However, hardly any technical work has attempted to implement virtue ethics in moral machines.[3] The main reason is that virtue ethics has been proven difficult to approach from a computational perspective, especially in comparison to its more popular alternatives. Simply put, it is easier to implement particular

✉ Jakob Stenseke
jakob.stenseke@fil.lu.se

[1] Department of Philosophy, Lund University, Lund, Sweden

[1] For simplicity, we will use terms such as "machine", "AI system", and "computer" interchangeably, referring to computational systems in both software and hardware.

[2] Cervantes et al. (2020) defines an AMA as: "a virtual agent (software) or physical agent (robot) capable of engaging in moral behavior or at least of avoiding immoral behavior. This moral behavior may be based on ethical theories such as teleological ethics, deontology, and virtue ethics, but not necessarily" (p. 506).

[3] In a comprehensive survey of more than 50 implementations in machine ethics, not a single one had virtue ethics as its main focus (Tolmeijer et al. 2020). Overall, only a handful of published articles explicitly deal with the construction of artificial virtuous agents (AVAs), and none of them provide any implementation details.

Published online: 15 December 2021

deontological rules and consequentialist utility-functions as opposed to generic virtues and moral character.

The main goal of this paper is to tackle the challenge head-on and demonstrate how virtue ethics can be taken all the way from theory to machine implementation. The rest of the paper is structured as follows. In Sect. 2, we explore four major benefits virtue ethics could offer to the prospect of moral machines. In Sect. 3, we face up to four critical challenges for artificial virtue, including (3.1) the uncodifiability of virtuous language, (3.2) its reliance on human-like moral capacities, (3.3) the role of virtues, moral exemplars, and eudaimonia, and (3.4) issues regarding technical implementation. Simultaneously, we outline a path to artificial virtuous agents (AVAs) based on moral functionalism, bottom-up learning, and eudaimonic reward. Section 4 describes how features of the outlined virtue-ethical framework can be interpreted in terms of functionality, which in turn can guide the technical development of AVAs. We then present a generic architecture that can act as a blueprint for algorithmic implementation. In Sect. 5, we discuss remaining challenges and identify promising directions for future work.

2 Machine ethics and virtue ethics

2.1 Machine ethics

With a growing number of self-driving vehicles on public roads, and a variety of robots being in education, medicine, and elderly care, it is hard to question the urgent need for AI systems to have some form of ethical considerations factored into their decision making. Were AI to continue on its course of replacing traditional human roles such as drivers, medical doctors, soldiers, and teachers, we should also expect them to adequately meet the moral standards entailed by such roles. For these reasons, machine ethics have attracted a growing amount of interest among academics in the intersection of moral philosophy and computer science, and the resulting body of work ranges from more or less detailed prototypes of ethical machines to theoretical essays on what moral agents ought or ought not to do (Anderson and Anderson 2011; Sparrow 2007; Winfield et al. 2014).

There are many pathways towards ethical machines, and different paths provide their own distinct benefits and disadvantages. The rule-based nature of deontological ethics elegantly corresponds to the type of conditional statements often associated with computer code. Similarly, the utility-maximizing aspects of consequentialism seem to resonate well with objective functions found in mathematical optimization, or the reward functions used in reinforcement learning. Deontology and consequentialism are thus fruitful frameworks for the pursuit of moral machines in their own ways, each corresponding to important aspects of moral behavior found in humans. However, by building AMAs based on these approaches, there is a risk of cherry-picking particular features of the theoretical counterpart without offering any account of how these are situated in the general cognition of the agent and its relation to the complicated dynamics of our every-day ethical lives. Behind rules and utility-functions, there is no moral character to speak of; no learning or adaptation; no thorough account of what it is to be moral besides following a simplified version of the normative theory it is based upon.

2.2 The appeal of virtue ethics

"Virtue ethics" refers to a broad family of related ethical views, with variations found in Buddhist, Hindu, and Confucian traditions (Flanagan 2015; Perrett and Pettigrove 2015; Yu 2013). In the Western tradition, the most influential version stems from Aristotle, and in the modern age it has emerged on to the central stage owing a lot to Anscombe (1958), Nussbaum (1988), Hursthouse (1999), and Annas (2011). While a comprehensive summary of the tradition and its variations could fill a large library,[4] we will introduce some of its central aspects by highlighting four major benefits the theory offers to moral machines.

2.2.1 Moral character

First and foremost, a person following virtue ethics puts her main focus on her character by fostering the dispositions that enable her to act in a morally good way. In this sense, virtues are the morally praiseworthy character traits one has or strives to possess. The courageous agent puts itself at risk to save another agent, not because it follows a rule, nor because it would result in the best outcome, but simply because it is what a courageous agent does (Hursthouse 1999). Consequently, the virtuous agent blurs as well as bridges the gap between, on the one hand, actions as a result of conscious deliberation and reasoning, and on the other hand, the psychological and biological dispositions that enable her to act in certain ways. While normative theories can be useful heuristics that guide us towards ethical conduct in terms of principles and reasons, in everyday life, there is often a gap between how we ideally ought to act and how we actually act.[5] In fact, our general behavior is influenced by a range of conscious and unconscious processes; from emotions and motivations at the psychological level to mood-altering

[4] For two excellent introductions to virtue ethics, see Crisp and Slote (1997) and Devettere (2002).

[5] To take a trivial example, we all know how hunger can affect our thinking and behavior in various ways: we get angry and frustrated without reason, or consume unhealthy food although we know that it is bad for us.

hormones and gut bacteria of our biological systems (Tangney et al. 2007; Teper et al. 2011). Taking character as the principal subject of moral evaluation, virtue ethics therefore enables us to account for these mechanisms, which in turn allows us to conceptualize a more comprehensive picture of what it means to be moral.

2.2.2 Learning from experience

Another key feature of virtue ethics is the emphasis on learning from experience.[6] A child and an adult might share the same good intentions, but due to a lack of experience, the child is often unaware of what she needs to do to effectively reach the intended result. Only through experience can we acquire the practical wisdom (phronesis) that helps us exercise good judgment and promote the excellence of our character and habits. Fueled by the same intuition, experimental studies in moral psychology have grown into full-blown research paradigms that seeks to illuminate the ways cognitive development and experience are necessary for certain forms of moral conduct.[7] In a similar vein, Annas (2011) has developed an influential version of virtue ethics based on the idea that the way we learn to be virtuous is similar to how we acquire a practical skill (Dreyfus 2004). Following Annas, moral competence is acquired both in terms of judgement (e.g., to follow reasons) and action (e.g., a moral know-how) through an active intelligent practice, akin to how we acquire and exercise skills such as farming or playing the piano.

One advantage of taking a learning approach to machine morality is that it enables the AMA to be context-sensitive to particulars and adapt to changes in ways that are difficult to encode in static rules. After all, real-life moral dilemmas rarely present themselves in the abstract and distilled manner as they are often portrayed in thought-experiments such as the trolley problem. It is perhaps even rarer that we find a moral dilemma to be in every sense similar to some we have encountered before, which in turn curtails the applicability of general principles. Following Aristotle, it is rather through our repeated encounter with particulars that we practice our practical wisdom (NE 1141b 15). Consequentially, as the virtuous agent develops through a continuous interaction with its environment, it would ideally be able to conform, not only to certain values and rules, but to the subtler details of social norms and cultural customs, with the additional ability to adapt accordingly as these change over time.

2.2.3 Relationship to connectionism

A third selling-point for the prospect of artificial virtue, and a natural extension of the second, is that virtue ethics resonates well with connectionism[8] and the correlated methods that are frequently employed in modern AI. Although ideas about artificial neural networks and learning algorithms have circulated since the 1940s (McCulloch and Pitts 1943), connectionism truly rose to prominence through the speed of twenty-first century computer chips combined with internet-age amounts of training data (Miikkulainen et al. 2019). Faster processing and more data have allowed increasingly larger networks, which in turn has made machine learning and neural networks the most dominant AI tools of today, reaching human and expert-level performance in areas such as pattern recognition, game playing, translation, and medical diagnosis (Deng and Yu 2014; Senior et al. 2020). Due to the learning emphasis, and the ability to capture context-sensitive information without the use of general rules, several authors have pointed out the appeal of uniting virtue ethics with connectionism (Berberich and Diepold 2018; Gips 1995; Howard and Muntean 2017; Wallach and Allen 2008). Some would even go so far as to claim that connectionism holds the essential keys to fully account for the development of moral cognition (Casebeer 2003; Churchland 1996; DeMoss 1998), while others have noticed the historical link between virtue ethics and connectionism through Aristotle.[9] Essentially, the relationship between the two could therefore provide AVAs with a compelling cognitive framework in combination with the technological backbone of modern learning methods.

2.2.4 Relationship to general cognition

The last appeal is that virtue ethics, compared to other theories, cuts deeper into the relationship between moral cognition and cognition in general. That is, virtue ethics situates morality, not separate from, but rather alongside general capacities and functionality. To use an analogy: we often measure the performance of artificial systems in functional terms, i.e., to the extent they are able to perform a certain task.[10] If they were equipped with more salient forms of moral behavior, we would also judge their behavior in relation to their other capacities. For instance, the moral

[6] In the words of Aristotle: "[…] a young man of practical wisdom cannot be found. The case is that such wisdom is concerned not only with universals but with particulars, which become familiar from experience" (NE 1141b 10).

[7] From the pioneering work by Piaget (1965) and Kohlberg and Hersh (1977), to modern refinements by Gilligan (1993) and Rest et al. (1999).

[8] Connectionism is the cognitive theory that mental phenomena can be explained in terms of artificial neural networks.

[9] For instance, Medler (1998) has pointed out that Aristotle was the first thinker to propose some of the fundamental concepts of connectionism.

[10] A good coffee machine is one that reliably produces great-tasting coffee. Given a set of symptoms, an apt medical diagnostics system is able to determine, with a high accuracy, what disease a person is likely to have.

competence of a self-driving car is intimately linked to its general ability to drive safely without supervision, including capacities such as speed control and collision-detection. A self-driving car with faulty brakes would simply lack the ability to avoid certain collisions even if the control-system of the vehicle was determined to do so. Similarly, you can only be courageous if you have the means to act courageously; to save a person from drowning in the ocean, you need to know how to swim. The point is that morality cannot be viewed isolated from non-moral capabilities. This reflects Plato and Aristotle's shared view on how virtue is intimately related to function; a virtue is a quality that enables you to be good at performing your function (*ergon*).[11] A good knife has the virtues—of being durable, sharp, etc.—that allows it to carry out its function (cutting). A good life, according to Aristotle, is thus to fulfill one's *ergon* through virtuous living (*arete*). In a modern context, this can be seen to emphasize the intimate relationship between moral cognition and general cognition. For instance, some authors have argued that there is no sharp distinction between moral and non-moral cognition on the basis that they have coevolved throughout the evolution of mankind (Flanagan 2009; Kitcher 2011).[12] Or as Johnson (2012) claims: there is no special moral faculty besides the general faculties. In the growing imaging literature of moral cognition, the emerging picture is that morality relies on a highly diverse and decentralized neural network that selectively uses specific regions depending on the associated context (FeldmanHall and Mobbs 2015).

Grounding morality in a general cognitive framework is constructive for the pursuit of moral machines in several regards. It allows us to more clearly determine what the appropriate virtues would be for an artificial agent in relation to its role, and help us to focus on the relevant traits that enable it to excellently carry out its function. A social companion robot used in elderly care should not share the same virtues as a self-driving car; they are equipped with different functionalities, serve different purposes, and face their own distinct problems. By contrast, the prejudice of universalist moral philosophy, i.e., the idea that there are general answers to particular moral problems, might lead one to implement one and the same "generic moral module" in machines across all domains, which would obstruct the nuances and domain-specific challenges that machines face with regards to their particular purpose.

Placing artificial morality within general cognition would also enable the development of AMAs to continuously draw from insights from the growing body of brain science, which in turn could shed light on aspects of morality that are only possible through the use of other complex and highly distributed cognitive abilities.

In summary, virtue ethics provides a smorgasbord of attractive features for the pursuit of moral machines. However, we have so far only explored the prospect of AVAs in rather idealistic terms. To construct an AVA that would fully realize the discussed benefits, one would need to create something more or less similar to a virtuous human being, which is unrealistic given today's technology.

3 Challenges for artificial virtue

Approaching virtue ethics from a computational perspective presents several novel challenges. In this section, we will focus on issues stemming from (i) the equivocal nature of the theory and its concepts, (ii) its reliance on human-like moral capacities, (iii) the difficulty of deciding the role of virtues, and (iv) technical implementation. Using the moral functionalism of Howard and Muntean (2017) and Hursthouse's virtue-ethical framework (1999), we will argue that there is a feasible path towards artificial virtue, but only if we give up the idea of trying to capture the full depth of the theory's anthropocentric roots.

3.1 The uncodifiability of virtuous language

The first challenge is to translate the concepts of virtue ethics into implementable computer models. This immediately becomes a difficult task since virtue ethics originated in ethical traditions through rich vocabularies of often interrelated, ambiguous, and higher-order mental concepts. In other words, the language of virtue ethics relies on thick descriptions, thick concepts,[13] and folk psychology.[14] To promote the traits that enable her to be courageous and fair, a virtuous person needs to have a thick understanding of courage and fairness to relate them to her own experience

[11] In 1.7 of the *Nicomachean Ethics*, Aristotle suggests that we can get at a clearer conception of *eudaimonia* (flourishing) if we first can determine the *ergon* (function or purpose) of human beings (NE 1.7 1097b 24). He justifies this inquiry by writing "for all things that have a function or activity, the good and the 'well' is thought to reside in the function" (NE 1.7 1097b 26–27). This echoes the famous knife-parable in Plato's *Republic*; that each thing has a function which can be identified by considering what the thing can achieve on its own or better than anything else (R 1.352e).

[12] In a similar vein, Casebeer has proposed that connectionism can serve as a suitable framework for a naturalized ethics, where moral capacities such as judgement and blame can be understood in a biological context among other abilities that allows an organism to skillfully deal with the demands of the environment (Casebeer 2003).

[13] While thick concepts are both "evaluative and descriptive" (Blackburn 1998), thick descriptions are those that embed subjective explanations and context into their meaning, e.g., describing an individual's behavior by extrapolating on its internal motivations (Geertz 1973).

[14] Folk psychology refers to the psychological common-sense ability to explain and understand mental states (Goldman 1993).

and motivations.[15] "Courage" and "fairness" are also paradigmatic examples of thick concepts, i.e., terms can be characterized descriptively while simultaneously having an evaluative quality.[16] Furthermore, since folk psychological notions such as belief, desire, and intention play a crucial role in our everyday ethical lives, to foster one's character arguably implies an ability to grasp the way such concepts are grounded in mental states of others and oneself (Dennett 1989).

However, no AI systems can apprehend rich contexts, nor can they make use of a catalogue of subjective experience; nor do they possess interpretative mechanisms to disentangle value-laden terms or follow the logics of commonsense psychology. To that end, virtue ethics can be hard to compile even for a human being. One common criticism of the theory is that virtue ethics is "uncodifiable" and does not offer an applicable decision-procedure (Hursthouse 1999, pp 39–42).[17] Faced with a particular moral dilemma, virtue ethics does not provide any straight-forward solutions; we simply have to trust that we do what a virtuous person would do.

The aforementioned difficulties have led many machine ethicists to avoid virtue ethics completely, others to argue that it is inferior to other approaches, or that it is simply incomputable due to its uncodifiability (Arkin 2007; Bauer 2020; Tolmeijer et al. 2020). But we will argue that there is a path for artificial virtue, provided that we give up on the project of trying to fully accommodate the theory's anthropocentric foundations.

A simple version of the incomputability-argument can be constructed on the assumption that machines are essentially systems of automated rule-following. Since machines are governed by rules, and a virtuous person is not, it follows that virtue ethics is incomputable. This line of reasoning, however, ignores the fact that AI systems can be constructed of rudimentary rule-adhering units while the behavior of the larger system is not rule-following in the same sense. The relevant analogy is found in the similarity between the neurons of biological minds and the nodes of an artificial neural network. Biological neurons receive and transmit impulses according to the all-or-none law, meaning that they either produce a maximum response or none at all.[18] Still, the human brain—consisting of roughly 86 billion neurons—is able to support complex processes that are not rule-following in the same way as its smallest components. After all, it is the very same network that gives rise to the thought and comprehension that enables a person to act virtuously. By extension, a large artificial neural network can produce a variety of behaviors that are not rule-adhering in the same narrow sense its nodes are.[19]

Howard and Muntean (2017) have extended a series of similar analogies between human and artificial cognition based on connectionism that they believe can pave the way for AMAs based on a form of virtue ethics. Drawing on Jackson and Pettit (1995) and Annas (2011), at the core of their framework is a "moral dispositional functionalism" which emphasizes "the role of the functional and behavioral nature of the moral agent" (Howard and Muntean 2017, p. 134). In their view, virtues are seen as dispositional traits that are nourished and refined through active learning of moral patterns in data, similar to how a cognitive system adapts to its environment (Howard and Muntean 2017). It is possible that Howard and Muntean's vision might in the long-term solve some of the challenges posed by virtuous language; through active exposure to particulars, the AVA eventually learns to approximate the functional role of generic virtues, and how they are related to each-other in a complex whole. The first step towards AVAs is therefore not to create an artificial human being, fully equipped with the abilities required to grasp virtue ethics in a "top-down" fashion. It is rather to construct "bottom-up" learners who continuously interact and adapt to a dynamic environment, and through experience develop the appropriate dispositions depending on their functional role. Albeit lacking a reference to virtue ethics, previous technical work has already explored learning-methods to tackle the ambiguity of moral language. Using neural networks, Guarini (2006, 2013a, b) have taken a "classification" approach to moral data with a focus on the gap between particularism and generalism. After learning, Guarini's models are able to classify cases as morally permissible or impermissible without the explicit

[15] Virtue ethics have found great success by drawing on 'thick' backdrops of culture and tradition, of shared experiences and stories; a leading example being the Christian tradition, which is full of accounts of virtuous living through the lens of individual experience. A similar use of thick descriptions can be found in virtue ethics' emphasis on moral exemplars. Indeed, many versions of virtue ethics stress the importance of observing and learning from moral role-models, from Aristotle (NE 1143b 1) and the Christian concept *Imitatio Christi* ("What would Jesus do?"), to prominent modern accounts (Hursthouse 1999; Zagzebski 2010).

[16] Thick concepts, and particularly the way they straddle the is-ought distinction, have been subject to long-standing metaethical debates. While some argue that the evaluative and non-evaluative aspects of thick concepts can be disentangled into separate components (Blackburn 1992; Hare 1991), others view them as inseparable fusions of fact and value (Putnam 2002; Williams 2006).

[17] However, some virtue ethicists welcome this feature on the basis that it is unrealistic to provide a straight-forward moral code based on virtue ethics (McDowell 1979).

[18] In other words, the neurons can be seen as following the simple rule "if excitation from stimuli is over a certain threshold" → "fire neuron".

[19] This does not, however, exclude the possibility for AI systems to be rule-following in the sense that they learn to follow rules or make use of a rule-following heuristics.

use of principles. Similarly, McLaren have developed two systems—*TruthTeller* and *Sirocco*—that can learn and reason from moral data with the purpose of supporting humans in ethical reasoning (McLaren 2005, 2006).[20]

Due to its inability to provide a decision-procedure, some still view virtue ethics as merely a supplement to the action-guidance provided by deontology and consequentialism. But others have found consolation in Hursthouse's virtue ethics since it both provides (a) a decision-procedure in terms of rules, and (b) accounts for the developmental aspects of morality (Hursthouse 1999). According to Hursthouse, virtue ethics can offer action guidance through rules that express the terms of virtues and vices ("v-rules"), such as "do what is courageous" or "do not what is unjust". While the list of virtues that yield positive rules of action is relatively short, it is complemented by a significant number of vices that can be expressed as negative rules (e.g., "don't be greedy"). With regards to a decision-procedure, Hursthouse writes "P.1. An action is right iff it is what a virtuous agent would characteristically (i.e. acting in character) do in the circumstances" (p. 28).[21]

Even if Hursthouse's decision-procedure is useful for the algorithmic implementation of virtue ethics, it might not be necessary as such from the view of system design. While Hursthouse's virtue ethics can offer action-guidance for humans who are conflicted about what they ought to do, the same conflict does not necessarily arise for AVAs with dispositional virtues. An AVA that serves as a lifeguard and saves someone from drowning does so, not because of the conscious deliberation of decisions they could have made, but because they *are* courageous. Even if the agent in question followed an algorithm that can be described as a decision-procedure, the central focus is not the procedure itself, but rather the way it enabled the agent to save the person in danger. In fact, the concept of a decision-procedure, and the presumed requirement of having one to implement a normative theory, are entrenched with assumptions of how human rationality works. It assumes that ethical behavior is conducted in a rather stepwise algorithmic fashion, which in turn disregard the role of affective dispositions and enforces the sort of "particular situation → particular action" analysis akin to deontology and consequentialism. Since virtue ethics seeks to unite both thinking (e.g., conscious deliberations that can take the form of a decision procedure) and feeling (e.g., attitudes, emotions, desires) under term "character", it should thus not be reduced to a description of the former.[22] Simply put, a character is not a decision-procedure.

Thus, instead of taking a "top–down" approach to virtue ethics through its thick concepts as seen from a human perspective, a productive path forward is to start from some functional interpretation of virtue ethics that would carry out at least some important aspects of the theory. To that end, we have outlined an approach to AVAs that emphasize a holistic conception of character (involving both non-affective deliberation and affective dispositions) and bottom-up learning using artificial neural networks.

3.2 Virtuous capacities: rationality, autonomy, and consciousness

The second set of challenges is in many ways a corollary of the first: to what extent do AVAs rely on "higher-order" forms of moral capacities? Lacking human-like rationality, subjective experience, and autonomy, one might question whether artificial agents can be attributed moral agency at all. While such concerns perturb the overall possibility for machine morality, we will focus on how it challenges the prospects of artificial virtue. On the basis that it is both unfeasible and ethically problematic to equip artificial agents with human-like morality, we will argue that the development of AVAs should instead be driven by functional capacities that are shaped by normative considerations of how and to what extent AI systems should be involved in human practices.

In the context of moral machines, there has been widespread debate regarding the sufficient and necessary conditions for moral agency.[23] Central to these discussions are rationality,[24] autonomy,[25] and consciousness[26] (from

[20] However, while Guarini and McLaren's systems show how learning methods can be applied to process moral data, they provide no further insight into the creation of virtuous agents; that is, agents capable of engaging in a variety of moral behavior beyond classification of moral data and text-retrieval.

[21] Furthermore, Hursthouse does not deny that 'thick' virtuous terms can be difficult to apply in any given situation, but instead stresses, following Aristotle and Plato, the importance of moral education; not merely through the indoctrination of rules, but the training of moral character (including motivation, emotion, and rationality).

[22] If such a decision-procedure was articulated it would need to include affective steps that seem rather absurd from the point of rational deliberation, such as "someone is in danger → "feel brave" → "save person".

[23] This includes debates on whether and to what machines can have a moral status, and whether and to what extent they should. For detailed discussions, see (Behdadi and Munthe 2020; Bryson 2010; Floridi and Sanders 2004; Gunkel 2014; Himma 2009; Johnson and Miller 2008; Sharkey 2017; Sparrow 2021; Tonkens 2009; Van Wynsberghe and Robbins 2019; Yampolskiy 2013).

[24] The role of rationality in the moral machine context has been explored in more detail by Coeckelbergh (2010); Hellström (2013); Himma (2009).

[25] For instance, while several authors claim that autonomy and free will are essential for moral agency (Hellström 2013; Himma 2009), others have argued that artificial free will is impossible (Bringsjord 2008; Shen 2011).

[26] The claim that phenomenal consciousness—i.e., the subjective "what it is like"—is necessary for moral agency has been advocated by authors such as Champagne and Tonkens (2015); Coeckelbergh (2010); Himma (2009); Linda (2010); Purves et al. (2015); Sparrow (2007).

now on collectively referred to as RAC). The good life for Aristotle's animale rationale is life lived in accord with reason, implying an ability to follow reason, e.g., for holding beliefs and performing actions. It also entails introspective capacities of conscious deliberation and rational inquiry. As such, it differs significantly from the mere "goal directed behavior" of rational agents as conceived in AI development (Russell and Norvig 2020). Perhaps even more central to morality is autonomy, as it forms a basis for discussions about free will, moral agency, and responsibility.[27] In the Kantian tradition, a person is autonomous only if her actions, choices, and self-imposed rules are without influence of factors inessential or external to herself (Kant 2008). However, such accounts of autonomy are very different from the functional autonomy found in AI systems, where it roughly refers to the ability of doing something independent from human control.[28] Furthermore, it seems difficult to have ideas of "good" and "bad" without the conscious experience of positively or negatively valenced states. But from a neuroscientific and computational point of view, consciousness remains more or less as elusive as it was when Descartes wrote *cogito ergo sum*.

Beyond long-standing metaphysical debates, e.g. between free will and determinism (autonomy), and body and mind (consciousness), the prospect of artificial RAC is also ethically problematic, as it might result in human suffering,[29] artificial suffering,[30] or artificial injustice.[31] There is also a danger that RAC and similar terms can be used to reproduce ideas that have been, and still are, used to justify abuse and hierarchies of dominance.[32]

It seems clear that we cannot, at least in a near-term, construct AMAs with human-like RAC,[33] and even if we could, we need a much deeper understanding of RAC to even properly assess whether we should. The present project is, however, not to create artificial humans, but to construct agents that are able to serve important roles in human practices. As pointed out by Coleman, while Aristotle's *human arete* emphasizes a life of contemplation and wisdom, the quest towards AVAs ought to be guided by exploring the *android arete* (Coleman 2001). For instance, it is possible to model rationality that allows artificial agents to effectively pursue goals without necessarily relying on the meta-cognitive abilities of human rationality. Besides self-legislative autonomy, there are flexible ways to construe artificial autonomy that enable human operators to oversee, intervene, or share the control of the system to avoid unwanted consequences.[34] Additionally, learning systems can employ simple reward functions that functionally mimic aspects of the role subjective preferences have in human cognition, without the phenomenological experience of suffering.[35]

More importantly, following the "normative approach" to artificial moral agency (Behdadi and Munthe 2020), we believe that the *android arete* should be guided and constrained by the normative discussions of how AI systems should engage in human practices that normally presuppose responsibility and moral agency. That is, rather than focusing on theoretical discussions on whether AI systems can have moral agency,[36] the development of AVAs should be led by the ethical and practical considerations that relate to their specific role, e.g., as doctor assistants, chauffeurs, or teachers. In turn, this allows us to shift focus from general questions about moral capacities based on human-like RAC, to particular issues about how and whether certain

[27] According to the "Principle of Alternative Possibilities", an agent can only be morally responsible for an action if she had the option to do otherwise (Frankfurt 1969).

[28] For instance, a helicopter that can fly from point A to B without a human controlling it is autonomous with regards to the very specific ability to fly from point A to B. Still, the helicopter was programmed to perform this very specific task; it is not autonomous in the sense that it set its own goal, wrote its own code, or had an option to say "no".

[29] The catastrophic risks of future AI have been extensively explored in Bostrom (2014); Russell (2019); Tegmark (2017).

[30] Metzinger has called for a global moratorium on synthetic phenomenology (i.e. artificial consciousness), as it could potentially lead to an explosion of artificial suffering (Metzinger 2021).

[31] Tonkens have argued that the very creation of an autonomous AMA based on virtue ethics is morally permissible by the tenets of the theory itself, e.g., forcing autonomous moral agents to perform tasks they have no choice to not perform violates virtues associated with social justice (Tonkens 2012).

[32] As pointed out by Cave (2020), rationality shares an intimate historical relationship with "intelligence", a value-laden term that have been widely used to legitimize domination; from Plato's "philosopher king", the logics of European colonialism and eugenics, to the present-day mass-slaughter of sentient non-human beings.

[33] In a recent review of the current status of AMAs, Cervantes et al. (2020) concludes that "there are no general artificial intelligence systems capable of making sophisticated moral decisions as humans do", and "… there is a long way to go (from a technological perspective) before this type of artificial agent can replace human judgment in difficult, surprising or ambiguous moral situations" (p. 527).

[34] For instance, adjustable autonomy is an active research area that has presented many fruitful ways to mitigate various issues related to autonomous systems, including ethical challenges and risks (Mostafa et al. 2019).

[35] In the long-term, if artificial phenomenology holds the key to more profound forms of excellent behavior, it might be possible to engineer capacities that are functionally equivalent to conscious experience but lack the "what it is like"-component (Besold et al. 2021).

[36] Behdadi and Munthe (2020) argues that much of the philosophical AMA debate is "conceptually confused and practically inert" (p. 195) on the basis that (i) supposedly essential features that drive the discussion (e.g. rationality and free will) can be understood in dispositional terms, (ii) it fails to include practical considerations such as responsibility allocation, and (iii) it is unclear how specific concepts of moral agency relate to actual artificial entities.

AI systems should be incorporated into specific ethical domains, what moral roles they could potentially excel at, and how agency and responsibility ought to be allocated in those circumstances.

Besides carrying out a role, it is also possible that the envisioned virtuous agents could still acquire a certain moral status. In the context of social robotics, Gamez et al. (2020) have argued that AVAs could claim membership to our moral community on the basis of two separate but consistent views on moral status: behaviorism (Danaher 2020) and the social-relational approach (Coeckelbergh 2010; Gunkel 2018). According to the former, an artificial agent has a moral status if they are functionally equivalent to other moral agents.[37] In the latter, the moral status of an artificial agent depends on the meaningful social relations we develop with it, such as reciprocal trust, duties, and responsibilities. In their experiment, Gamez et al. (2020) found that while individuals made weaker moral attributions to AIs in comparison to humans, they were still willing to view the AIs as having a moral character. Thus, even if AVAs lack the unique metaphysical qualities of human morality, if we are willing to describe them as having a character—based on their behavior and our relationship to them—it could be sufficient reason to welcome them to our moral community.

3.3 Virtues, moral exemplars and eudaimonia

The third challenge is to decide the role of virtues in the moral cognition and behavior of AVAs. From Homer to Benjamin Franklin, many different lists of virtues have seen the light of day, emphasizing different aspects of ethical life (MacIntyre 2013). Some lists have been more prominent than others, in particular the cardinal virtues; prudence, justice, fortitude and temperance.[38] This might suggest that one could feed an artificial virtuous agent with widely accepted virtues, or generic virtues suitable for machines.[39] However, this solution would only be an option if there was (i) a universally agreed-upon list of the most essential virtues, and (ii) a way to implement said list in a top-down fashion. As argued in 3.1, even if a list was attainable, the approach to virtue has to be bottom-up since the only way to reach context-sensitive generals are through particulars. We therefore agree with MacIntyre's historical analysis, that virtues ought to be based in a particular time and place, emerging out of the community in which they are to be practiced (MacIntyre 2013).

Still, this leads us to the question: in what way should a virtuous agent learn bottom-up? Inspired by Hursthouse (1999) and Zagzebski (2010), previous work in artificial virtue have centered on imitation learning through the role of moral exemplars (Berberich and Diepold 2018; Govindarajulu et al. 2019). The moral exemplar approach offers several appeals. By mimicking excellent virtuous humans, we do not have to worry about what virtues they in fact end up with since they would replicate something that is already virtuous. Besides providing means of supervision and control, imitation learning would also solve the alignment problem; i.e., the challenge of aligning machine values with human values (Berberich and Diepold 2018).

Nevertheless, there are issues with the moral exemplar-focus; in particular the challenge of deciding who is a moral exemplar and why. After all, there could be severe disagreements about who is and who is not a virtuous person. According to Zagzebski (2010), exemplars can be recognized through the emotion of admiration, which allows us to map the semantic extension of moral terms to features of moral exemplars. Govindarajulu et al. (2019) have provided a rudimentary formalization of Zagzebski's suggestion using deontic cognitive event calculus (DCEC). In their model, admiration is understood as "approving (of) someone else's praiseworthy action" (p. 33), which depends on a primitive emotional notion of pleased or displeased based on whether an action led to some positive or negative utility. Using the utility of consequences to define emotions, however, their model seems to be driven by consequentialism rather than virtue ethics.[40]

Besides moral exemplars, we suggest that there is an alternative source for moral evaluation appropriate for bottom-up virtuous agents to be found in the concept of *eudaimonia* (conventionally translated as "well-being" or "flourishing"). Instead of relying on moral exemplars or a list of anthropocentric virtues, Coleman (2001) has argued for an eudaimonist approach to artificial virtue, where "all of one's actions aim at a single end—in Aristotle's case, happiness (eudaimonia)—and virtues are those character traits which

[37] The use of behaviorism as a foundation to evaluate the moral performance of machines have given raise to debates regarding the possibility and credibility of a moral Turing test (Arnold and Scheutz 2016; Gerdes and Øhrstrøm 2015).

[38] Initially derived from Plato (R, 4.426–435), accepted by the Stoics, and later appropriated and developed within Christian theology by thinkers such as Thomas Aquinas and Augustine of Hippo (Bejczy 2011).

[39] For instance, Berberich and Diepold (2018) have argued that equipping AVAs with the virtue temperance would solve the control problem; i.e., the issue of making sure that some future superintelligent AI would not harm or destroy its creators (Bostrom 2014). Their point is that if an AVA had temperance it would not seek "limitless self-improvement" and thus not pose an existential threat to humans (Berberich and Diepold 2018).

[40] Whereas Zagzebski's view finds intuitive support in the way virtues and moral terms are exemplified in human narratives of heroes and saints, Govindarajulu et al.'s model rather presents a formal way in which artificial agents can admire excellent consequentialist exemplars.

foster the achievement of this end" (p. 249). Besides avoiding circular definitions of virtue (e.g., "virtues are qualities of virtuous individuals"), eudaimonia can explain the nature of virtues in terms of a goal or value to strive towards.[41]

An eudaimonist virtue ethics offers several benefits for the prospect of AVAs. Essentially, it enables us to model virtues in terms of their relationship to eudaimonia. If eudaimonia is defined as "increase moral good X", virtues would then be the traits that help the agent to increase X. In machine learning terms, eudaimonia can be seen as the reward function that informs the learning and refinement of virtues and virtuous action. In this way, the artificial agent will become virtuous in the sense that it develops the dispositions that enable it to effectively pursue a certain goal or increase a certain value (depending on whether eudaimonia is defined as a goal or value). Another strength is that, while learning through imitation is limited to mere behavior, an eudaimonist approach can encompass values both intrinsic (e.g., hedonistic pleasure and pain) and extrinsic (e.g., values that support human ends).[42] Furthermore, in cases where it is hard to settle on a suitable moral exemplar, a functional eudaimonia offers a "top-down backdoor" to implement certain values or goals that are then attained through a bottom-up learning process.

But adopting a eudaimonist view raises the further question: what should be the eudaimonia of virtuous agents? While the content of eudaimonia can be defined as the goal that ought to be achieved, or the good that ought to be increased, we believe that the important function of eudaimonia is that it provides a moral direction for the virtuous agent; a measure to evaluate and refine its moral character and virtues. However, this omits the difficult task of pinning down the actual goal or values an AVA should have.[43] But there are good reasons to remain cautiously silent on the de facto content of an AVA's eudaimonia. First, it allows us to use eudaimonia as a functional placeholder for a wide variety of values and ends, on the premise that they can be implemented in computational systems (which we will discuss in Sect. 4.3). For practical reasons, it might be suitable for different AVAs to have different types of eudaimonic content depending on their functional role. Second, it recognizes the ambiguity of human eudaimonia, especially as it remains unclear whether and to what extent artificial systems can apprehend the complexity of the former.[44] This echoes Hursthouse (1999), who views eudaimonia as a value-laden concept that is intentionally ambiguous to allow for interpretative headroom and disagreement.[45]

3.4 Technical implementation

The remaining challenge is to move from the conceptual plane towards technical implementation. We do so by examining previous work in artificial virtue, focusing on what it can fruitfully provide for the development of AVAs.[46]

Howard and Muntean (2017) have put together a web of conceptual foundations for the construction of artificial autonomous moral agents (AAMAs). Beyond moral functionalism and bottom-up learning, they conjecture a number of, in their words "incomplete and idealized analogies" (p. 137) between human cognition and machine learning that can guide the deployment of AAMAs through a combination of neural networks and evolutionary computing. However, the actual details of the implementation are missing, and only partial results of an experiment are provided where neural networks have learned to detect irregularities in moral data.[47] The biggest flaw with their project, however, is that

[41] This neo-Aristostotelian version of virtue ethics also finds support in Hursthouse (1999). Although, Hursthouse's view has also been somewhat misused in the machine context as a reason to solely focus on moral exemplars (Gamez et al. 2020; Tonkens 2012), in particular the first premise for virtue theoretical action-guidance (Hursthouse 1999, p. 28).

[42] Of course, this opens up the question whether the AVAs should be seen as mere tools for certain human-defined ends, or whether they should pursuit their own ends. Both views are potentially problematic. To make sure that artificial eudaimonia align with human values, we believe the first should be the present focus, but only given the premise that the AVAs are not subject to suffering nor that we violate their moral autonomy (Metzinger 2021; Tonkens 2012). If they, on the other hand, defined and pursued their own ends, it opens up a potential path to unforeseeable risks (Bostrom 2014).

[43] After all, it is the same grand challenge the consequentialist faces when she tries to define moral goods, and roughly similar to the challenge a deontologist faces when deciding upon the right duty. For instance, the problem has divided consequentialism into several camps, including hedonistic pleasure (Bentham and Mill), satisfaction of preferences (Singer 2011), rule utilitarianism (Hooker 2002), combination of actions and rules (Hare 1981), state welfare (Mozi), and the reduction of suffering (Smart 1958).

[44] Human eudaimonia can potentially involve a large set of different goals and values that are spatially and temporally distributed. For instance, our everyday happiness can depend on the success of our local football team, our current health, possibility to achieve long-term career goals, ability to do the things we enjoy, or some broader ideas of global well-fare (Hursthouse 1999). However, no current artificial system can comprehend an abstract phrase such as "well-being of all sentient beings", nor could it simply try to reduce suffering in the universe.

[45] For an interesting description of how eudaimonia can be construed from conflicting views, see Hursthouse (1999, pp. 188–189).

[46] We omit the already discussed work (Coleman 2001; Gamez et al. 2020; Govindarajulu et al. 2019; Tonkens 2012).

[47] Another issue with their work is that they seem to conflate the concept of moral autonomy with some form of independence achieved through the stochastic trial-and-error of evolutionary computation methods, writing "evolution of populations of neural networks endows the AAMA with more moral autonomy…" (Howard and Muntean 2017, p. 152), and that successive generations are "gradually more independent" (p. 152). Perhaps one could say that the AAMAs autonomously found irregularities in patterns, but only in the sense that they were trained to do so independently, and not in the sense that they were morally autonomous (as discussed in 3.2).

it is practically infeasible,[48] and furthermore, it is not clear how their envisioned agents should be deployed in moral situations besides classification tasks explored by Guarini (2006).

In a similar vein, Berberich and Diepold (2018) have explored the technical underpinnings of artificial virtue based on connectionist methods. Most interestingly, they have described how reinforcement learning can be used to shape the moral reward function of virtuous agents in three ways: (i) through external feedback from the environment, (ii) internal feedback by means of self-reflection, and (iii) observation of moral exemplars. However, besides providing a broad outline of potential AVA features, they offer no finer details of how such an agent could be constructed.

Thornton et al. (2016) have incorporated principles from virtue ethics along with deontology and consequentialism into the design of automated vehicle control. In their hybrid model, deontology determines vehicle goals in terms of constraints, consequentialism in terms of costs, and virtue ethics is used to regulate the strength of the applied costs and rules depending on the vehicle's "role morality". "Role morality" refers to behaviors that are acceptable given the context of a particular professional setting.[49] For instance, it is acceptable for an ambulance to break traffic laws—e.g. by running a red light—if it transports a passenger with life-threatening conditions. Essentially, Thornton et al. (2016) shows how the moral character of an AI system can be defined with regards to their societal role, and how the character can be modeled using "virtue weights" that balance costs and constraints that enables them to perform their function.

4 Artificial virtuous agents

Based on the takeaways from our investigation so far, we now turn to the task of interpreting the outlined theory in terms of functionality, which in turn can guide the further development of AVAs. The aim is not to provide a detailed implementation per se, but rather to discuss suitable methods that can be combined to functionally carry out features of virtue ethics in a variety of moral environments. Essentially, the viability of the proposed framework rests on the assumptions that (i) the function of dispositional virtues can be carried out by artificial neural networks (Sect. 4.2), (ii) eudaimonia can be functionally interpreted as a reward function that drives the training of the virtue networks (Sect. 4.3), and (iii) modern learning methods can in various ways support the development of artificial phronesis (Sect. 4.4).[50]

4.1 Artificial character

In virtue ethics, a moral character can be defined as the sum of an agent's moral dispositions and habits (George 2017). In our functional model, it consists of several components: a set of stable yet dynamic dispositions (virtues), a reward function (eudaimonia), a learning system (phronesis), and relevant mechanisms for perception and action determined by its role (e.g. input-sensors, memory, and locomotion). The moral character thus denotes the entire character, encompassing both moral qualities and non-moral qualities that enable it to perform its role. Virtues are "stable yet dynamic" in the sense that they are fixed at a given moment but have the ability change over time. Importantly, instead of applying a decision-procedure to a particular situation, the virtuous character continuously interacts within an environment based on its internal states.

4.2 Artificial virtues and vices

Given our eudaimonist view, virtues are defined as the character traits an agent needs and nourishes to function well in light of its eudaimonia. We extend the weight analogy of Thornton et al. (2016) and the classification approach of Guarini (2006) and suggest that the functional aspect of virtues can be captured in the function of nodes in an artificial neural network. The essential role of virtues in this view is that they, based on some input from the environment, determine the action taken by the agent (e.g., its output).

[48] Their main idea is to evolve entire populations of neural networks through an evolutionary algorithm that changes the topology, parameter values, and learning functions of the networks. Through fitness selection, the emerging AAMA is the one with "a minimal and optimal set of virtues that solves a large enough number of problems, by optimizing each of them" (p. 153). While certain combinations of randomized search methods and neural networks have yielded promising results in limited applications through deep reinforcement learning and NEATs (Berner et al. 2019; Stanley and Miikkulainen 2002), Howard and Muntean's suggested project turns into a search problem of infinite dimensionality, that in effect requires infinite computational resources. Additionally, due to the highly stochastic and open-ended process of evolving artificial neural network, it is not at all obvious that excellent moral agents would emerge even given infinite computational resources.

[49] Following Radtke (2008), moral roles are based on societal expectations derived "from a collective decision on what is best for society" (Thornton et al. 2016, p. 1436).

[50] More broadly, the relationship between the proposed artificial virtuous cognition and human cognition relies on connectionism (as discussed in Sect. 2.2). For more in-depth examinations of this relationship, see Howard and Muntean (2017), Casebeer (2003), and DeMoss (1998).

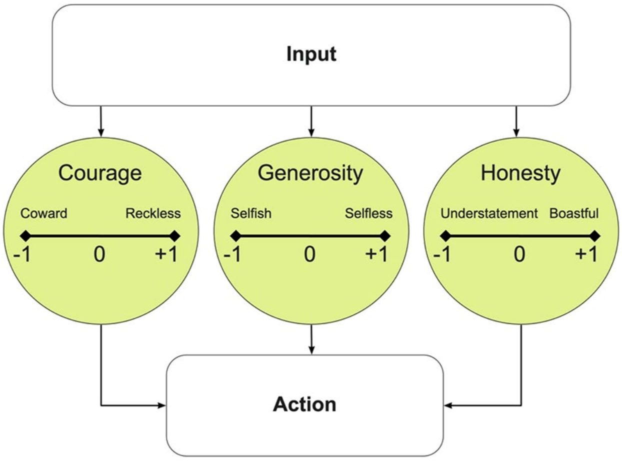

Fig. 1 Illustration of a network with three virtuous perceptrons. Three types of inputs are parsed into three corresponding virtues weights, each holding a value between two extremes (in this case represented numerically from −1 to+1). Each perceptron produces one out of two possible actions depending on its weight. For instance, if another agent is in danger, an agent with a positive courage weight (>0) will try to help the other agent even if it poses a risk

In the simplest case, a virtue can be modeled as a perceptron that determines whether an agent acts in one way or another given a certain input. A perceptron is a threshold function that takes an input x and produces the output value $f(x) = 1$ if $w \times x + b > 0$, where w is the weight (or set of weights) and b is the bias. By either increasing or decreasing the weight through feedback, the perceptron is effectively a learning algorithm for binary classification. In a system of perceptrons, the nodes themselves represent virtues since there is only one neural unit for every virtue (as illustrated in Fig. 1). This solution is suitable if there are only two possible actions (e.g. "save" or "don't save"), and the binary virtues need to find an appropriate balance between two distinct vices (e.g. "courage" as a balance between "reckless" and "coward").[51]

In more complex applications, a virtue can consist of a larger network of nodes that receive an input and output one of many possible actions. In this view, virtues can be seen as a "higher level" amalgamation that encompasses a large set of particular "lower level" units (see Fig. 2). More nodes enhance the ability to process more detailed information that can in turn be used to produce more fine-tuned actions.

Deciding on what virtues to implement and the actions they ought to perform entirely depends on context and functionality of the AVA in question. It also informs the choice between static and dynamic virtues and weights, e.g., whether one implements the virtues and weights one prima facie believes to be suitable for the AVA, or let them learn independently in light of an eudaimonic reward (discussed in Sect. 4.3). In an environment with a fixed set of possible actions but a wide range of environmental inputs, it would be suitable to provide the agent with static virtues relating to the fixed set of actions, but with dynamic weights in order for them to learn the appropriate action given a specific input. In an environment where we already know that a particular input always ought to be followed by a particular action, it would be more appropriate to provide the virtue with a static weight.[52] However, in highly dynamic and noisy environments with a potentially infinite number of possible actions, the agent might be limited to unsupervised learning and reinforcement learning in accord with a reward function.[53]

In addition to the choice between static or dynamic, it is also important to consider how the system deals with the conflict problem, i.e., the issue that arises when two or more virtues suggest different actions.[54] For instance, in a social situation, compassion might tell us to remain silent while honesty urges us to convey some painful truth. One solution is to resolve conflicts through mere comparison of strength, i.e., given a particular situation, the right action depends on what the most dominant virtue tells the agent to do. If an agent is more fair than selfish, it will still give food to the begging other. In a computational setting, such conflicts can be resolved by simple arithmetic; if fairness weight = 0.6 and selfishness weight = 0.4, then fairness > selfishness. A related solution is to model virtues in a pre-given hierarchy of priorities. Another option is to train the input-parsing network of the virtuous agent so it learns to map inputs to the appropriate virtue-network. This could be achieved through a process of supervised learning, where the network is presented with many pre-labeled scenarios that are related to specific virtue-networks.[55] Yet another option is to model more sophisticated forms of hybrid virtues that can combine and parse different aspects of inputs that relate to different virtues. Essentially, as there are many practical ways to combine and resolve conflict between virtues, we do not believe the conflict problem raises any serious concerns for the prospect of AVAs.

[51] Note the resemblance to the virtue-theoretical concept of a "golden mean"; as a fine-tuned balance between two extremes.

[52] In this sense, the virtuous agent inherits the rule-following robustness of a deontological agents.

[53] Learning methods will be discussed in more detail in Sect. 4.4.

[54] Note that the conflict problem also targets deontological agents in the case of conflicting rules, or consequentialist agents in the case of conflicting utilities.

[55] It would therefore be similar to how an object recognition network can predict the most likely object to feature in an image or video.

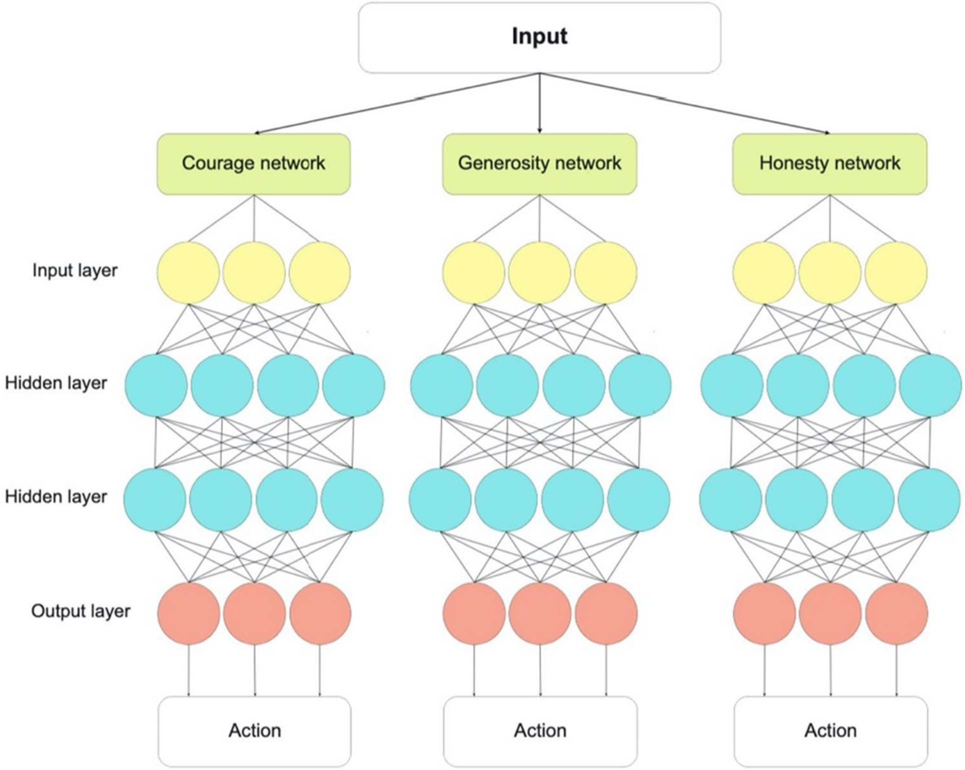

Fig. 2 Illustration of a system with three virtuous networks, each with an input layer, two hidden layers, and an output layer. Compared to the perceptron, a deep neural network can deal with classification tasks that are not linearly separable

4.3 Artificial eudaimonia

Approaching artificial eudaimonia, the main challenge is to provide AVAs with a definition of eudaimonia that can be practically implemented in computational systems. Towards a possible solution, we propose a functional distinction between eudaimonic type (e type) and eudaimonic value (e value). We define e type as the kind of value or goal the virtuous agent strives towards (i.e., the eudaimonic content as described in Sect. 3.3), and *e* value as the quantitative measure of how much e-type that has been attained. Importantly, *e* type and *e* value provide the basis for learning, as the dispositional virtues change in light of whether an action increased or decreased the *e* value. To be functionally implementable, an e type must represent a preference to increase *e* value given by some identifiable measure (e.g., a quality or quantity), and that *e* value can increase or decrease through the agent's actions. An e type can for instance be a preference to increase praise and decrease blame received from others. In that case, the agent needs to have some way of receiving feedback on their actions, and to qualitatively recognize the feedback as either praise or blame. With an e type defined by praise/blame-feedback, the agent's *e* value will then increase if it receives praise and decrease if it receives blame. Another way of illustrating a functional e type is through resources. We can imagine that there is some quantifiable resource that agents need to survive (e.g. food). A selfish e type could then be defined as a preference to maximize the possession of said resource, and a selfless e type could conversely be defined as a preference to give resources to others. The selfish agent then increases its *e* value through actions that increase the resource (e.g., begging or stealing), and the selfless agent increases *e* value by giving.

One might question whether and to what extent our rather simplistic conception of eudaimonia can, in some meaningful sense, capture the variety and complexity of human values. In particular, our proposal rests on the rather strong assumption that moral goods and goals can be described in quantifiable measures. Furthermore, since we leave the choice of e type to human developers, it also raises the question whether the implemented values will be justified in relation to the AVA's role.[56] While we do not have any definite answers to these issues, we believe that parts of the solution lie in further developments and experimental studies, and that our model provides a starting point for such endeavors. Even if no current AI system can apprehend the full depth of human values, it does not exclude the possibility for there to be some moral domain where some quantifiable moral good can be legitimately increased through computational means.

[56] As discussed by Tonkens (2012), this in turn raises important questions such as "who is a virtuous engineer?", and whether virtuous engineers can develop artificial agents more virtuous than the engineers themselves.

4.4 Artificial phronesis

While *phronesis* ("practical wisdom") is a rather ambiguous concept within the virtue theoretic tradition,[57] in our functional simplification, we take artificial phronesis to broadly refer to the learning an agent receives from experience. This interpretation is motivated by recognizing the central role learning plays in cognition, for moral and non-moral capacities alike (Annas 2011). We will address four learning aspects of artificial phronesis, namely what is learned, how it is learned, the source of learning, and the technical method used.[58]

For what and how, we make a distinction between (1) learning what action leads to good and (2) learning what is good in itself.[59] Based on our conception of artificial eudaimonia, (1) can be seen as the instrumental means to increase e-value according to some e-type, whereas (2) refers to an ability to change and refine the teleological component itself (e-type). We identify three features of (1) that we believe are crucial for the development of artificial phronesis and describe how they can be acquired: (1a) virtuous action, (1b) understanding of situation, and (1c) understanding of outcome.

(1a) Virtuous action refers to the ability to perform virtuous action, i.e., to act courageously or fairly. Given the action-determining role of virtues in our model, the key element in the development of virtuous action is to fine-tune virtues in light of eudaimonic feedback. However, to do so, the agent needs (1b) understanding of situation and (1c) outcome.

(1b) Understanding of situation refers to the ability to comprehend a situation, i.e., to know what input relates to what virtue; for instance, whether a situation calls for courage or honesty. (1c) Understanding of outcome, on the other hand, is the ability to comprehend what was the actual result of a performed action. While virtuous action can be trained by means of reinforcement using eudaimonic feedback, we suggest that (1b) can be carried out by an input parsing network trained on labeled data of various moral situations, i.e., by means of supervised learning.[60] Similarly, (1c) can consist of a network trained to recognize an outcome in light of the relevant quantity or quality at stake as defined by its e type, e.g., by learning from a dataset of labeled reactions. For instance, in a simple environment driven by praise/blame reactions, the role of the outcome network is to accurately classify a reaction as being either "praise" or "blame".

The combined role of these features can be illustrated in the following example. An agent receives input from the environment in terms of another agent in need of help. The input parsing network classifies the situation as involving courage, and therefore sends the input to the courage network. In turn, the courage network assesses the risk of the situation at hand and determines whether it calls for a certain action (e.g., depending on the network's balance between recklessness and cowardice). The outcome of the performed action is then read by the outcome network as a new input, and parses it to the eudaimonic reward system, which evaluates whether the action increased or decreased *e* value. If *e* value increased, positive reinforcement is sent to the courage network, in effect teaching the courage network that the performed action was appropriate. By contrast, negative reinforcement will reduce the likelihood that the same action will be performed given a similar scenario in the future.

Beyond trial-and-error (through external feedback) and supervised learning (provided by developers), AVAs could also learn from internally generated feedback (Berberich and Diepold 2018). We describe two possible routes that could guide the construction of an internal learning system, namely retrospective and proactive reflection. Given that an agent has the ability to store mappings between input-action-outcome-reward, it could analyze that information to retrospectively reinforce certain actions. Identifying patterns in past behavior, retrospective reflection could in turn form the basis of more nuanced behavior in complex environments based on statistical considerations (e.g., by applying non-linear regression). Proactive reflection, on the other hand, could be achieved through internal simulation of possible scenarios, where learning feedback is based on trial-and-error of hypothetical input-virtue-action-outcome-reward mappings.[61]

[57] For Plato, all forms of virtuousness are a form of phronesis (Guthrie 1990). For Aristotle, phronesis is not only necessary to achieve a certain end, but for living well in general. Additionally, phronesis has more recently been equated with the virtue of prudence, meaning the ability of "seeing ahead" or "foresight". In the words of Vallor (2016), practical wisdom unites "perceptual, affective, and motor capacities in refined and fluid expressions of moral excellence that respond appropriately and intelligently to the ethical calls of particular situations" (p. 99).

[58] Note that terms such as "know" or "understand" are used in this section as functional concepts that we assume can be applied to artificial systems. For instance, the term "know" can be potentially confusing in this context, since artificial agents with dispositional virtues do not "know" in the epistemic sense of the term. It is therefore closer to Aristotle's concept of techne (practice or "know how") as opposed to episteme (theoretical knowledge).

[59] The distinction is based on the difference between instrumental good (i.e., means to some other good) and intrinsic good (good in itself).

[60] This is similar to how an object recognition network can predict the most likely object to feature in an image or video.

[61] In deep reinforcement learning, learning through "competitive self-play" can produce behaviors more complex than the training environment itself (Bansal et al. 2017), and reach superhuman performance in complex game environments such as Go or Dota 2 (Berner et al. 2019; Silver et al. 2018). This could inspire a potential path for "self-play" in simulated moral environments where an artificial agent

Another potential source for learning is the behavior and experiences of others, regardless of whether they are exemplars or not.[62] If an AVA can observe that the outcome of another agent's action increased some identifiable moral good (defined by the observer's e type), it could teach the observer to positively reinforce the same action.[63]

Evolution offers yet another potential source for learning. This could be achieved through the use of evolutionary computation (Howard and Muntean 2017), or other randomized search methods. The main idea behind evolutionary algorithms is to find candidate solutions to optimization problems using processes inspired by biological reproduction and mutation, along with a fitness function that evaluates the quality of the solutions (Bäck et al. 1997). A possible application to the development of artificial phronesis could therefore be to (i) generate an initial population of agents with virtues and other suitable parameters set randomly, (ii) evaluate the fitness of every individual according to how much e-value they have attained, (iii) select the most virtuous individuals for reproduction and generate offspring using crossover and mutation, (iv) replace the least virtuous individuals with the new offspring, and repeat steps (ii)–(iv) until the population is sufficiently virtuous.

We have thus far only been concerned with (1) learning what action leads to good as opposed to (2) learning what *is* good in itself. The latter can be achieved by modelling dynamic e-types that have the ability to change. There is an intuitive appeal for such an endeavor, as the ability to change our (human) concept of eudaimonia provides an important basis for personal, social, and moral progress. Yet, it also presents a puzzling paradox—how can we, on the basis of our current set of values, assess whether another set of values are more appropriate?[64] Beyond such meta-theoretical issues, there are good reasons why dynamic e-types in the context of moral machines should be approached with caution, as it can potentially pave the way for nonalignment of human-AI values[65] and reward hacking.[66]

Still, we believe that there are a few suitable venues to explore dynamic e-types. The first is through the use of moral exemplars (discussed in 4.5), and the second is by means of a metaheuristic at system level. In a multi-agent environment, there can be some potential "higher good" to be achieved at a system level that cannot simply be resolved at the level of individuals. For instance, in a "tragedy of the commons" situation, individuals act in their own self-interest even though the collective action of the many creates catastrophic problems for everyone (such as a systemic collapse). Using randomized search methods in a multi-agent simulation of virtuous agents, dynamic e-types could then be used to identify e-types that satisfy hedonistic needs of individuals while simultaneously ensuring the prosperity of the entire population at large.[67]

Finally, we will briefly discuss technological methods that can be used to develop artificial phronesis. Methods in machine learning are conventionally divided into supervised, unsupervised, and reinforcement learning, each offering their own unique set of advantages and drawbacks (Russell and Norvig 2020). In the first, a function learns to map the correct input to output based on labeled training data; in the second, it learns to categorize and find patterns in unlabeled data on its own; in the third, an agent learns to make actions that maximize some cumulative reward through feedback. Agreeing with Berberich and Diepold (2018), we believe that reinforcement learning (RL) provides the most appealing approach for AVAs, as it, contrary to the other two, is based on dynamic interaction with an environment and thus supports a continuous process of learning from experience (so called "online learning"). We have already described how RL constitutes the basis for the eudaimonic reward system, where e type corresponds to the reward function of the RL agent, and *e* value is the measure of how much reward

Footnote 61 (continued)

continuously interacts with other versions of itself and tries out a vast number of different actions to produce novel forms of morally excellent behavior.

[62] Intuitively, we learn a lot from observing others' right- and wrongdoings even if we do not share the same virtuous traits or even the same idea of eudaimonia. For instance, the concept of an "immoral" or "evil" person, and the social repercussions of being perceived as one, can potentially be a powerful source for moral development (Haybron 2002).

[63] Correspondingly, if the moral good decreased, the agent could learn to avoid the action.

[64] To illustrate, we can imagine Ewa, who leads a life in search of fame and fortune, solely seeking pleasure in material wealth and recognition. At some point Ewa encounters a wise elder telling her that what she seeks is not what she "truly wants and needs". At first, the wise elder is wrong, since fame and fortune are in fact what Ewa truly wants given her current conception of eudaimonia. However, as the wise elder explains how the quest for fame and fortune comes at the expense of other essential features of well-being—ignoring love and friendship—Ewa becomes increasingly convinced that what she currently seeks won't bring her true happiness. She therefore changes her

Footnote 64 (continued)

old idea of eudaimonia and starts to cultivate other virtues that will propel her towards the new.

[65] Dystopian science-fiction abounds with examples showing why the development of conscious AI, pursuing their own ideas of happiness, is a very bad idea. If an artificial virtuous agent were set "in the wild", free to explore different concepts of eudaimonia, it could, as in Bostrom's example, in the end turn the whole universe into paperclips (Bostrom 2020).

[66] Reward hacking is what occurs when reinforcement learning agents find a way to maximize their reward in a manner that conflicts with the developer's original intentions (Amodei et al. 2016).

[67] A similar project is to study the evolution of cooperation among self-interested agents, as done extensively in game theoretical models following the work of Axelrod and Hamilton (1981).

is attained. With that said, supervised and unsupervised methods can also be fruitfully integrated into our model. In particular, we have described how supervised learning can be used to train the input parsing and outcome networks of AVAs. Unsupervised learning such as cluster analysis could also find suitable applications in the form of anomaly detection, for instance, by helping the agent to identify deviations and outliers in the internally stored moral data.

4.5 Moral exemplars

To incorporate moral exemplars in our framework, we have to provide some good solutions to two challenges, namely (i) how to pick a suitable moral exemplar, and (ii) how to learn from them. From a practical point of view, and given the limited capacities of current AMAs (Cervantes et al. 2020), the most obvious moral exemplar for AVAs is the human exemplar. While the use of human exemplars raises its own set of issues (see Sect. 3.3), it is nevertheless the human developer who defines and implements the e-types, learning system, virtues, and evaluates agent performance. But we will briefly outline a possible approach to moral exemplars that can be useful in computational settings derived from our model. The conditions are the following:

One agent (X) takes another agent (Y) as a moral exemplar if

i. X and Y have the same e type, and
ii. Y has a higher *e* value than X.

The first condition means that X and Y strive towards the same set of values and goals. The second means that Y has been more successful in achieving the same set of values and goals, e.g., due to its virtuous behavior.[68] Since X wants to achieve the same thing as Y (same e type) and recognizes that Y has some means of achieving it more effectively (more *e* value), it is reasonable for X to take Y as a moral exemplar. Although these conditions are difficult to model in the interaction between AVAs and humans (it would require that humans have a formally defined e type), we believe it can be implemented in the interaction between different AVAs, (e.g., in multi-agent simulations).

Adopting the outlined approach to artificial moral exemplars, the second issue—regarding how an agent should learn from exemplars – depends on the information provided by the environment. If agents only have access to the external behavior of others, we are limited to learning through behavioral imitation. However, in that case it might be impossible for agents to determine whether they should adopt a moral exemplar at all, since they would not have access to the e type or *e* value of others.[69] Alternatively, if all internal aspects of AVAs were accessible, agents could adopt moral exemplars by simply copying relevant aspects of their character, e.g., the structure and weights of the virtue networks. If they, on the other hand, had access to the e type and *e* value of others but not the virtue networks as such, it would still provide sufficient reason, given the conditions outlined earlier, to adopt exemplars and learn from them through behavioral imitation.

4.6 Software architecture

The final challenge is to explicate the appropriate connections between the different components of the AVA.[70] Ideally, the connections should be drawn in a way that effectively exploits component functionality while leaving room for learning through continuous exploration. We believe that the functionality of the discussed features informs such a design, and conclude by a stepwise explanation of the core aspects of a generic architecture that can guide the development of AVAs (Fig. 3):

1. *Input parsing network*: input from the environment is classified by the input parsing network. Its main role is to transmit the environmental input to the appropriate virtue network. It corresponds to understanding of situation, i.e., to know what virtue applies to a particular situation. The network can be trained through supervised learning using labeled datasets of various ethical scenarios. One critical aspect of the network is to determine whether a situation calls for moral action or not,[71] and if

[68] However, this does not exclude the possibility that Y has achieved a higher *e* value due to moral luck (Williams 1981). A potential solution to the problem of moral luck is to define additional conditions of the type (iii) "Y has not, to the best knowledge of X, achieved a higher *e* value merely due to circumstances", or (iv) "Y has achieved a sufficiently high *e* value to the extent that it is statistically improbable that it was due to moral luck".

[69] Berberich and Diepold (2018) has suggested a solution to this problem by means of inverse reinforcement learning, where agents learn to imitate the behavior of moral exemplars by approximating their reward function (Ng and Russell 2000). However, their view of artificial eudaimonia is different than the one presented in our model; it is shaped by a complex relationship to virtues such as prudence, temperance, gentleness, and friendship. Simply put, in our model the function of e type is to shape the virtues, but in their view, the virtues shape the eudaimonic reward function.

[70] As proposed by Howard and Muntean (2017), a population of networks with an ability to change their topology could in principle develop the appropriate connections through randomized search methods. However, the practical infeasibility of such a project is more clear if we consider the sheer number of relevant connections needed to arrive at a functional agent.

[71] A similar capacity for "moral attention" is explored by Berberich and Diepold (2018).

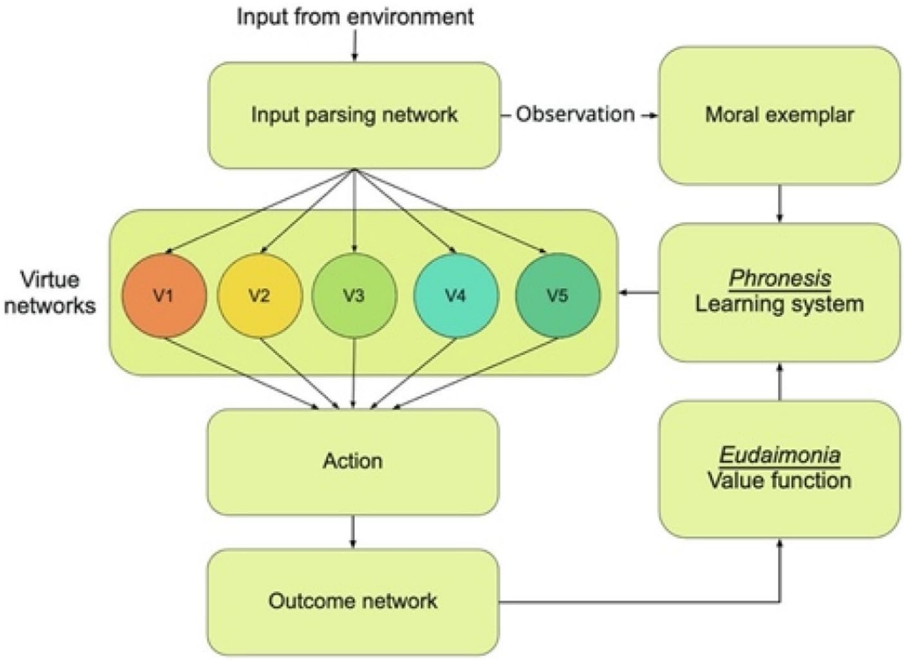

Fig. 3 Generic architecture of the proposed AVA. Input from the environment is classified by the input parsing network and sent to the appropriate virtue network (in the figure represented as V1–V5). Each virtue network represents a character trait that can produce a range of different actions. The outcome of the performed action is taken as a new input by the outcome network and evaluated by the eudaemonic reward function. This in turn informs the learning system whether a particular action of a particular virtue is to be positively or negatively reinforced

the latter, whether it can constitute a basis for learning through observation.

2. *Virtue networks*: the invoked virtue network classifies the input to determine the most appropriate action. The number of nodes and layers depend on environmental complexity and the number of possible actions. If the input can be linearly separable into one out of two actions, a single perceptron can carry out the classification, but in complicated cases, a deep neural network would be more suitable.
3. *Action output*: the action determined by the virtue network is executed by the agent. This could in principle be any type of action; from simple movements and communicative acts to longer sequences of skillful action.
4. *Outcome network:* new environmental input is classified by the outcome network, corresponding to understanding of outcome. Similar to the input parsing network, it can learn from labeled data of situation-outcomes, in particular in light of the relevant moral goods as defined by its e-type.
5. *Eudaimonic reward:* the eudaimonic reward function evaluates the classified outcome according to e type and current *e* value. If the *e* value is increased, it sends positive feedback to the learning system, and negative feedback if *e* value decreased.
6. *Phonetic learning system:* the learning system reinforces the relevant virtue according to the eudaimonic feedback. Another possible application of reinforcement learning is to send feedback to the input network. For instance, if the negative feedback was exceptionally high, it might suggest that the input parsing network transmitted the environmental input to an inappropriate virtue network; if exceptionally positive, the input parsing can be positively reinforced.
7. *Observation and moral exemplars:* two additional sources for learning can be implemented in the form of (i) observation of others and (ii) moral exemplars. In the first case, if another agent's action increased or decreased some identifiable e type, the observing agent could positively or negatively reinforce the same action.[72] In the second case, if an agent adopts a moral exemplar (given conditions described in 4.5), it could learn from it by either copying aspects of its character or by mimicking its behavior.

5 Discussion

We have described how AVAs can be constructed in a way that functionally carries out a number of core features of its theoretical counterpart, including virtuous action, learning from experience, and the pursuit of eudaimonia. We believe the development of both simpler and more advanced virtuous systems can be guided by the presented framework. In a minimal case, the training of an AVA can solely rely on the feedback of its own actions. More advanced agents could potentially learn from passively observing the behavior of

[72] As noted earlier, this rests on the assumption that the observer can recognize the action taken and have the means of performing the same it in the future.

other agents and moral exemplars, or by means of internal feedback systems (e.g. retrospective and proactive reflection).

Bauer (2020) has argued that a two-level utilitarian approach to AMAs is superior to a virtue-theoretic approach since it encompasses the essential features of the latter while realizing additional benefits. He gives four reasons to support his claim: (1) conditional rules can serve the function of dispositional traits, (2) utilitarian AMAs avoid reference to the intangible concept of 'virtue' that obfuscates the design of AMAs, (3) utilitarian AMAs can be pre-loaded with widely agreed-upon rules, such as human rights and legal codes, and (4) since utilitarian AMAs would follow moral rules, they would be ethically better than "typical human behavior" whereas AVAs, being modeled on human behavior, would not. We believe our work shows that Bauer is misguided on all four points. Responding to (1), the sole purpose of dispositional traits is not to yield rule-following behavior as such, but rather to produce moral behavior in context-sensitive situations where simple rule-following principles are not applicable. Answering (2), while we agree that "virtue" is in many ways an intangible concept, we have shown that it can be given a functional definition within our eudaimonic framework. Against (3), we believe that connectionist learning offers the most effective methods to implement generic rules into AI systems so that they are carried out appropriately. Furthermore, a functional e-type allows for top-down implementations of widely agreed-upon values, provided that such values can be formalized. Responding to (4), we agree that imitation learning from human exemplars is not sufficient, which served as reason to adopt a teleological approach to machine ethics.

Given how eudaimonic reward trains virtues in light of outcomes, one could argue that the presented model is simply "consequentialism with an extra layer".[73] But that would miss the point. Although teleological virtue ethics relies on some definition of moral good, it emphasizes the learning and dispositional aspects of how certain goods could in fact be increased. That is, while a form of consequentialism drives learning, it is the character's dispositional virtues that produce the actions.

More generally, we do not believe that a virtue-theoretic approach is superior to deontology or consequentialism in every regard, but rather that it draws our attention to important aspects of morality that are overlooked in the field of machine ethics. After all, an artificial entity would only be truly virtuous if it could follow moral rules and be sensitive to the consequences of its actions. Ultimately, we believe that a hybrid-approach to machine ethics is most suitable as it could potentially realize the benefits of the three grand theories. However, since virtue ethics digs deeper into what it *is* to be a moral agent, we believe it offers a sketch of what a hybrid system could look like. If AMAs were ever to possess a character or belong to our moral community, they would indeed share many aspects with an agent as perceived through the virtue-theoretic lens.

Nevertheless, a number of issues remain to be solved before we see the introduction of morally excellent AVAs in our everyday lives. Conceptual work is needed to resolve the conflicts between anthropocentric notions of morality and the formalization of such concepts in AI development, and technical work is needed to bring AMAs into the complex ethical environments of the real-world. For the prospect of AVAs presented in this work, one critical issue is the lack of explainability in neural networks, often referred to as the black box problem.[74] Although it is unclear whether the issue can ever be completely resolved, we believe it can be approached carefully.[75]

So what kind of AVAs can and ought to be constructed using our framework? We will outline some potential applications that can serve as venues for future work in the short-, mid-, and long-term.

For now, the current stage of artificial virtue is prototypical and confined to well-defined software environments and limited robotic tasks. Following the classification approach (Guarini 2006), AVAs could be trained to solve a range of moral classification tasks, including action selection, situation reading, and outcome understanding. For instance, one interesting venue for future work is to explore technical solutions to the conflict problem (i.e., when two or more virtues suggest different actions). Another application is to implement AVAs in multi-agent systems to study cooperation among self-interested individuals[76] or other forms of complex social behavior.[77]

[73] This is somewhat similar to our own criticism leveled against Govindarajulu et al. (2019) in 3.3.

[74] That is, while neural networks are universal function approximators, one cannot simply understand the structure of the approximated function by looking into the network (Olden and Jackson 2002). The inner workings of an AVA with very large neural networks could potentially be as opaque as the human brain. In the worst case, an AVA might behave immorally, but due the complexity of its network, we cannot understand why.

[75] The black box problem is an active research area in AI safety, and there are already some promising solutions (Cammarata et al. 2020).

[76] According to Danielson (2002), artificial agents achieve morality when they can"constrain their own actions for the sake of benefits shared with others" (p. 196). Similarly, Leben (2018) claims that morality adapted"in response to the problem of enforcing cooperative behavior among self-interested organisms" (p. 5). For an application of deep reinforcement learning in the study of altruism, see Wang et al. (2018).

[77] To that end, we believe AVAs can provide useful insights in the paradigm of social simulation (Edmonds and Meyer 2015).

In the mid-term, virtue ethics may be incorporated into a number of AI systems in real-world domains, particularly in complex environments where bottom-up learning offers the only route to moral sensitivity. Additional benefits can be achieved by integrating virtue ethics in human-inspired architectures (Cervantes et al. 2016). In that way, artificial virtue can be equally propelled by advancements in brain science as from new AI methods. Further developments in the mid-term could potentially explore more sophisticated models and methods for learning, reasoning, communication, social cognition, and computational autonomy, provided that such capacities are desirable and ethically justified in relation to the AVA's specific role (Behdadi and Munthe 2020).

In the long-term, AVAs might, as argued by Gamez et al. (2020), be legitimate members of our moral community. Most optimistically, AVAs might not only be morally excellent, but become moral exemplars to humans by conveying forms of morality that are yet to be discovered in our everyday moral landscape. However, as we discussed in Sect. 3.2, such projects should be approached with utmost caution and care; the development of increasingly more sophisticated AI systems in the moral domain walks a risky path of potentially causing an explosion of suffering, for human and artificial beings alike.

6 Conclusion

We have broadly explored various philosophical and technical dimensions of virtue ethics and developed a comprehensive framework for the construction of artificial virtuous agents based on functionalism, bottom-up learning, and eudaimonia. To our knowledge, it is the first work that presents a roadmap to artificial virtue that is conceptually thorough yet technically feasible. Ultimately, we believe that it offers a promising path towards excellent moral machines and hope that our work will inspire further developments of artificial virtue.

Acknowledgements This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation. The author is grateful to his colleagues at the Department of Philosophy and the Department Cognitive Science at Lund University for insightful discussions and feedback on previous versions of the paper, in particular Frits Gåvertsson for useful input on virtue ethics, Amandus Krantz and Christian Balkenius for technical advice, and Björn Petersson, Ylva von Gerber, and Jiwon Kim for reviewing an early draft of the paper.

Author contribution JS conceptualization, methodology, formal analysis and investigation, writing—original draft, writing—review and editing.

Funding Open access funding provided by Lund University. This work was partially supported by the Wallenberg AI, Autonomous Systems and Software Program—Humanities and Society (WASP-HS) funded by the Marianne and Marcus Wallenberg Foundation and the Marcus and Amalia Wallenberg Foundation.

Availability of data and materials Not applicable.

Code availability Not applicable.

Declarations

Conflict of interest The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

Abel D, MacGlashan J, Littman ML (2016) Reinforcement learning as a framework for ethical decision making. In: AAAI Workshop: AI, Ethics, and Society, 2016. Phoenix, AZ, p 02

Amodei D, Olah C, Steinhardt J, Christiano P, Schulman J, Mané D (2016) Concrete problems in AI safety arXiv preprint https://arXiv.org/160606565

Anderson M, Anderson SL (2008) ETHEL: TOWARD a principled ethical eldercare system. In: AAAI Fall Symposium: AI in Eldercare: New Solutions to Old Problems, p 02

Anderson M, Anderson SL (2011) Machine ethics. Cambridge University Press

Annas J (2011) Intelligent virtue. Oxford University Press

Anscombe GEM (1958) Modern moral philosophy. Philosophy 33:1–19

Arkin RC (2007) Governing lethal behavior: embedding ethics in a hybrid deliberative/hybrid robot architecture. Report GIT-GVU-07-11, Georgia Institute of Technology's GVU, Atlanta

Arnold T, Scheutz M (2016) Against the moral Turing test: accountable design and the moral reasoning of autonomous systems. Ethics Inf Technol 18:103–115

Axelrod R, Hamilton WD (1981) The evolution of cooperation. Science 211:1390–1396

Bäck T, Fogel DB, Michalewicz Z (1997) Handbook of evolutionary computation. Release 97:B1

Bansal T, Pachocki J, Sidor S, Sutskever I, Mordatch I (2017) Emergent complexity via multi-agent competition. arXiv preprint https://arXiv.org/171003748

Bauer WA (2020) Virtuous vs utilitarian artificial moral agents. AI Soc 35:263–271

Behdadi D, Munthe C (2020) A normative approach to artificial moral agency. Mind Mach 30:195–218

Bejczy IP (2011) The cardinal virtues in the middle ages: a study in moral thought from the fourth to the fourteenth century, vol 202. Brill

Berberich N, Diepold K (2018) The virtuous machine-old ethics for new technology? arXiv preprint https://arXiv.org/180610322

Berner C et al. (2019) Dota 2 with large scale deep reinforcement learning arXiv preprint https://arXiv.org/191206680

Besold TR, Zaadnoordijk L, Vernon D (2021) Feeling functional: a formal account of artificial phenomenology. J Artif Intell Conscious 8:147–160

Blackburn S (1992) Through thick and thin. In: Proceedings of the Aristotelian Society, vol suppl, pp 284–299

Blackburn S (1998) Ruling passions. Oxford University Press, Oxford

Bostrom N (2014) Superintelligence: paths, dangers. Oxford University Press, Strategies

Bostrom N (2020) Ethical issues in advanced artificial intelligence. Routledge

Bringsjord S (2008) Ethical robots: the future can heed us. AI Soc 22:539–550. https://doi.org/10.1007/s00146-007-0090-9

Bryson JJ (2010) Robots should be slaves close engagements with artificial companions: key social, psychological, ethical and design issues 8:63–74

Cammarata N, Carter S, Goh G, Olah C, Petrov M, Schubert L (2020) Thread: circuits. Distill 5:e24

Casebeer WD (2003) Moral cognition and its neural constituents. Nat Rev Neurosci 4:840–846

Cave S (2020) The problem with intelligence: its value-laden history and the future of AI. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 29–35

Cervantes J-A, Rodríguez L-F, López S, Ramos F, Robles F (2016) Autonomous agents and ethical decision-making. Cogn Comput 8:278–296

Cervantes J-A, López S, Rodríguez L-F, Cervantes S, Cervantes F, Ramos F (2020) Artificial moral agents: a survey of the current status. Sci Eng Ethics 26:501–532

Champagne M, Tonkens R (2015) Bridging the responsibility gap in automated warfare. Philos Technol 28:125–137. https://doi.org/10.1007/s13347-013-0138-3

Churchland PS (1996) Feeling reasons. Neurobiology of decision-making. Springer, pp 181–199

Coeckelbergh M (2010) Moral appearances: emotions, robots, and human morality. Ethics Inform Technol 12:235–241. https://doi.org/10.1007/s10676-010-9221-y

Coleman KG (2001) Android arete: toward a virtue ethic for computational agents. Ethics Inf Technol 3:247–265

Crisp R, Slote MA (1997) Virtue ethics. Oxford University Press, Oxford

Danaher J (2020) Welcoming robots into the moral circle: a defence of ethical behaviourism. Sci Eng Ethics 26:2023–2049

Danielson P (2002) Artificial morality: virtuous robots for virtual games. Routledge

Dehghani M, Tomai E, Forbus KD, Klenk M (2008) An integrated reasoning approach to moral decision-making. In: AAAI, pp 1280–1286

DeMoss D (1998) Aristotle, connectionism, and the morally excellent brain. In: The Paideia Archive: twentieth World Congress of Philosophy, pp 13–20

Deng L, Yu D (2014) Deep learning: methods and applications. Found Trends Signal Process 7:197–387

Dennett DC (1989) The intentional stance. MIT Press

Devettere RJ (2002) Introduction to virtue ethics: insights of the ancient Greeks. Georgetown University Press

Dreyfus SE (2004) The five-stage model of adult skill acquisition. Bull Sci Technol Soc 24:177–181

Edmonds B, Meyer R (2015) Simulating social complexity. Springer

Feldmanhall O, Mobbs D (2015) A neural network for moral decision making. In: Toga AW, Lieberman MD (eds) Brain mapping: an encyclopedic reference. Elsevier, Oxford

Flanagan O (2009) The really hard problem: meaning in a material world. MIT Press

Flanagan O (2015) It takes a metaphysics: raising virtuous buddhists. Snow 2015:171–196

Floridi L, Cowls J (2019) A unified framework of five principles for AI in society Issue 11, Summer 2019 1

Floridi L, Sanders JW (2004) On the morality of artificial agents. Mind Mach 14:349–379. https://doi.org/10.1023/B:MIND.0000035461.63578.9d

Frankfurt HG (1969) Alternate possibilities and moral responsibility. J Philos 66:829–839. https://doi.org/10.2307/2023833

Gamez P, Shank DB, Arnold C, North M (2020) Artificial virtue: the machine question and perceptions of moral character in artificial moral agents. AI Soc 35:795–809

Geertz C (1973) The interpretation of cultures, vol 5019. Basic books

George MI (2017) What moral character is and is not. Linacre Quar 84:261–274

Gerdes A, Øhrstrøm P (2015) Issues in robot ethics seen through the lens of a moral Turing test. J Inform Commun Ethics Soc 13:98–109

Gilligan C (1993) In a different voice: psychological theory and women's development. Harvard University Press

Gips J (1995) Towards the Ethical Robot. In Ford K, Glymour C, Hayes P (ed) Android Epistemology, MIT Press, Cambridge MA, p 243–252

Goldman AI (1993) The psychology of folk. Psychol Behav Brain Sci 16:15–28

Govindarajulu NS, Bringsjord S, Ghosh R, Sarathy V (2019) Toward the engineering of virtuous machines. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. pp 29–35

Guarini M (2006) Particularism and the classification and reclassification of moral cases. IEEE Intell Syst 21:22–28

Guarini M (2013a) Case classification, similarities, spaces of reasons, and coherences. In: Coherence: Insights from Philosophy, Jurisprudence and Artificial Intelligence. Springer, pp 187–201

Guarini M (2013b) Moral case classification and the nonlocality of reasons. Topoi 32:267–289

Gunkel DJ (2014) A vindication of the rights of machines. Philos Technol 27:113–132. https://doi.org/10.1007/s13347-013-0121-z

Gunkel DJ (2018) Robot rights. MIT Press, London

Guthrie WKC (1990) A history of Greek philosophy: Aristotle: an encounter, vol 6. Cambridge University Press

Hagendorff T (2020) The ethics of AI ethics: an evaluation of guidelines. Mind Mach 30:99–120

Hare RM (1981) Moral thinking: its levels, method, and point. Clarendon Press, Oxford; Oxford University Press, New York

Hare RM (1991) The language of morals, vol 77. Oxford Paperbacks, Oxford

Haybron DM (2002) Moral monsters and saints. Monist 85:260–284

Hellström T (2013) On the moral responsibility of military robots. Ethics Inf Technol 15:99–107. https://doi.org/10.1007/s10676-012-9301-2

Himma KE (2009) Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics Inf Technol 11:19–29

Hooker B (2002) Ideal code, real world: a rule-consequentialist theory of morality. Oxford University Press

Howard D, Muntean I (2017) Artificial moral cognition: moral functionalism and autonomous moral agency. Philosophy and computing. Springer, pp 121–159

Hursthouse R (1999) On virtue ethics. OUP Oxford

Jackson F, Pettit P (1995) Moral functionalism and moral motivation. Philos Quar 45:20–40

Johnson M (2012) There is no moral faculty. Philos Psychol 25:409–432

Johnson DG, Miller KW (2008) Un-making artificial moral agents. Ethics Inf Technol 10:123–133. https://doi.org/10.1007/s10676-008-9174-6

Kant I (2008) Groundwork for the metaphysics of morals. Yale University Press

Kitcher P (2011) The ethical project. Harvard University Press

Kohlberg L, Hersh RH (1977) Moral development: a review of the theory. Theory Pract 16:53–59

Leben D (2018) Ethics for robots: how to design a moral algorithm. Routledge

Linda J (2010) The functional morality of robots. Int J Technoeth (IJT) 1:65–73. https://doi.org/10.4018/jte.2010100105

MacIntyre A (2013) After virtue. A&C Black

McCulloch WS, Pitts W (1943) A logical calculus of the ideas immanent in nervous activity. Bull Math Biophys 5:115–133

McDowell J (1979) Virtue and reason. Monist 62:331–350

McLaren B (2005) Lessons in machine ethics from the perspective of two computational models of ethical reasoning. In: 2005 AAAI Fall Symposium on Machine Ethics

McLaren BM (2006) Computational models of ethical reasoning: challenges, initial steps, and future directions. IEEE Intell Syst 21:29–37

Medler DA (1998) A brief history of connectionism Neural. Comput Surv 1:18–72

Metzinger T (2021) Artificial suffering: an argument for a global moratorium on synthetic phenomenology. J Artif Intell Conscious 8:43–66

Miikkulainen R et al (2019) Evolving deep neural networks. Artificial intelligence in the age of neural networks and brain computing. Elsevier, pp 293–312

Mittelstadt B (2019) Principles alone cannot guarantee ethical AI. Nat Mach Intell 1:501–507

Mostafa SA, Ahmad MS, Mustapha A (2019) Adjustable autonomy: a systematic literature review. Artif Intell Rev 51:149–186

Ng AY, Russell SJ (2000) Algorithms for inverse reinforcement learning. In: Icml, p 2

Nussbaum MC (1988) Non-relative virtues: an Aristotelian approach. Midwest Stud Philos 13:32–53

Olden JD, Jackson DA (2002) Illuminating the “black box”: a randomization approach for understanding variable contributions in artificial neural networks. Ecol Model 154:135–150

Perrett RW, Pettigrove G (2015) Hindu virtue ethics. The Routledge companion to virtue ethics. Routledge, pp 75–86

Piaget J (1965) The moral development. Free Press, New York, p 1

Purves D, Jenkins R, Strawser BJ (2015) Autonomous machines, moral judgment, and acting for the right reasons. Ethical Theory Moral Pract 18:851–872. https://doi.org/10.1007/s10677-015-9563-y

Putnam H (2002) The collapse of the fact/value dichotomy and other essays. Harvard University Press

Radtke RR (2008) Role morality in the accounting profession—how do we compare to physicians and attorneys? J Bus Ethics 79:279–297

Rest JR, Narvaez D, Thoma SJ, Bebeau MJ (1999) DIT2: devising and testing a revised instrument of moral judgment. J Educ Psychol 91:644

Russell S, Norvig P (2020) Artificial Intelligence: A Modern Introduction, 4th edn, Pearson. http://aima.cs.berkeley.edu/newchap00.pdf

Russell S (2019) Human compatible: artificial intelligence and the problem of control. Penguin

Senior AW et al (2020) Improved protein structure prediction using potentials from deep learning. Nature 577:706–710

Sharkey A (2017) Can robots be responsible moral agents? And why should we care? Connect Sci 29:210–216

Shen S (2011) The curious case of human-robot morality. Paper presented at the Proceedings of the 6th international conference on Human-robot interaction, Lausanne, Switzerland,

Silver D et al (2018) A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362:1140–1144

Singer P (2011) Practical ethics. Cambridge University Press

Smart RN (1958) Negative utilitarianism. Mind 67:542–543

Sparrow R (2007) Killer robots. J Appl Philos 24:62–77

Sparrow R (2021) Why machines cannot be moral. AI Soc 36:1–9

Stanley KO, Miikkulainen R (2002) Efficient evolution of neural network topologies. In: Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No. 02TH8600). IEEE, pp 1757–1762

Tangney JP, Stuewig J, Mashek DJ (2007) Moral emotions and moral behavior. Annu Rev Psychol 58:345–372

Tegmark M (2017) Life 3.0: being human in the age of artificial intelligence. Knopf

Teper R, Inzlicht M, Page-Gould E (2011) Are we more moral than we think? Exploring the role of affect in moral behavior and moral forecasting. Psychol Sci 22:553–558

Thornton SM, Pan S, Erlien SM, Gerdes JC (2016) Incorporating ethical considerations into automated vehicle control. IEEE Trans Intell Transp Syst 18:1429–1439

Tolmeijer S, Kneer M, Sarasua C, Christen M, Bernstein A (2020) Implementations in machine ethics: a survey. ACM Comput Surv (CSUR) 53:1–38

Tonkens R (2009) A challenge for machine ethics. Mind Mach 19:421

Tonkens R (2012) Out of character: on the creation of virtuous machines. Ethics Inf Technol 14:137–149

Vallor S (2016) Technology and the virtues: a philosophical guide to a future worth wanting. Oxford University Press

Van Wynsberghe A, Robbins S (2014) Ethicist as designer: a pragmatic approach to ethics in the lab. Sci Eng Ethics 20:947–961

Van Wynsberghe A, Robbins S (2019) Critiquing the reasons for making artificial moral agents. Sci Eng Ethics 25:719–735

Wallach W, Allen C (2008) Moral machines: Teaching robots right from wrong. Oxford University Press

Wang JX, Hughes E, Fernando C, Czarnecki WM, Duéñez-Guzmán EA, Leibo JZ (2018) Evolving intrinsic motivations for altruistic behavior arXiv preprint https://arXiv.org/181105931

Williams B (1981) Moral luck: philosophical papers 1973–1980. Cambridge University Press

Williams B (2006) Ethics and the limits of philosophy. Routledge

Winfield AF, Blum C, Liu W (2014) Towards an ethical robot: internal models, consequences and ethical action selection. Conference towards autonomous robotic systems. Springer, pp 85–96

Yampolskiy RV (2013) Artificial intelligence safety engineering: why machine ethics is a wrong approach. Philosophy and theory of artificial intelligence. Springer, pp 389–396

Yu J (2013) The ethics of confucius and Aristotle: mirrors of virtue, vol 7. Routledge

Zagzebski L (2010) Exemplarist virtue theory. Metaphilosophy 41:41–57