Virtue ethics in autonomous agents

Tomasz Zurek (t.a.zurek@uva.nl)
Informatics Institute, University of Amsterdam, Science Park 900, Amsterdam, 1098XH

Dorota Stachura-Zurek (dorota.stachura1@gmail.com)
Independent researcher
The Netherlands

Abstract

The paper presents a model and a discussion of the computational representation of virtue ethics in autonomous devices. One of the key problems in formal modeling of virtue ethics is the computational representation of the concept of virtue. In our model, the virtue is represented by a set of minimal extents to which a set of values, relevant to the virtue, should be satisfied. A device will be moral if any decision made satisfies all relevant values above the declared thresholds.

Keywords: virtue ethics; autonomous devices; model

Introduction

The rapid development of AI-driven autonomous devices has raised a number of questions concerning the ethical context of decisions made by these devices. Although the existence of robot vacuum cleaners does not seem to pose serious risks, a number of more complex devices prove they may be dangerous.

The moral control over the behavior of autonomous devices remains enormously important in the areas where the results of decisions can be extremely harmful and the moral dilemmas are particularly serious (for example, the military domain). Furthermore, in our view, in addition to such devices making ethical decisions, they should not only be able to provide explanation of why their decisions are ethical (ex-post), but should also operate transparently: the moral evaluation of a given decision should be predictable (ex-ante). This requirement is closely related to the so-called value-alignment problem: how to ensure that AI systems are properly aligned with human values (Russell, 2019; Allen & Wallach, 2012). We believe that the transparency of the decision-making mechanism would likely ensure that the decision is made with respect to required moral values.

Although a number of computational approaches exist which model ethical theories, most of them focus on various kinds of deontological and consequentialist ethics (see (Tolmeijer, Kneer, Sarasua, Christen, & Bernstein, 2021) for a detailed analysis). Some researchers though have proposed a thoughtful insight into the applicability of the virtue-based paradigm by exploring how virtue ethics theories could be linked with autonomous devices, scrutinizing the teleological character of virtue ethics and analyzing the value alignment problem in this context (Berberich & Diepold, 2018).

In this paper we present a formal model of virtue ethics as well as a mechanism of ethical decision-making and discuss it in the light of the existing literature and the problem of disruptive technologies (like military autonomous devices). The formal model presented in the paper is an extended version of the model initially introduced in (Zurek & Stachura-Zurek, 2021).

Virtue ethics

The theories labeled as virtue ethics rely on the centrality of the concept of virtue. This and two other concepts fundamental to virtue ethics include (Hursthouse & Pettigrove, 2018):

- arête - that is virtue, understood as aptitude / disposition / an excellent trait of character, which makes a person good;
- phronesis - practical wisdom; the capacity to reason about virtues; a meta-virtue on which virtuous decisions depend (Burbules, 2019);
- eudaimonia - the end of human existence; a state of real, true happiness, human flourishing.

As opposed to focusing on the overall good consequences, or trying to determine what is wrong and what is right, virtue ethics is about what kind of people we want to be and how we wish to shape our lives. The fulfilment of the human potential, full development of oneself as a person - eudaimonia - can be seen both as a result of virtuous life, but also as leading a virtuous life as such (Saja, 2015). Modern virtue ethics include a number of variations inspired by Aristotle, Plato, the Stoics, Aquinas, but also philosophers like Confucius (Sim, 2007) (Yu, 2007), or Hume and Nietzsche (Swanton, 2015). The understanding of virtue ethics for the purpose of implementation in a decision making mechanism in this paper will be rooted in a broadly eudaimonist ethical framework, with the crucial concepts seen as follows: (1) virtues are the foundation for decision making; (2) phronesis consists in the reasoning mechanism; (3) eudaimonia is achievable. Given that we perform our research in the spirit of weak AI, applying *human* ethics to artificial agents, it follows that we expect them to *behave* in a certain sort of manner, so that the results would be advantageous to humans. Virtue ethics clearly focuses on the agent as a person; the objective of such research would not be then to create a virtuous machine imitation of a human in pursuit of its happiness, but a device that is able to make a decision which would be considered virtuous by

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

a human. We hardly entertain the idea of eudaimonia as a *machine's* happiness or flourishing; it can obviously only be applied to humans, and thus in this case would involve ensuring happiness/flourishing for humans *through* the decisions of an autonomous device.

Basic assumptions

Prior to the introduction of our model, a set assumptions needs to be presented:

Firstly, it is crucial to point out that we do not intend to examine any kind of *ethics of artificial moral agents*, including the moral agency or responsibility of machines; our approach embraces machine ethics understood as an attempt to develop an ethical framework for moral and explainable decision-making in machines. Our intention is then not to create a framework for "robot morality," but to simulate the process of decision-making which takes into consideration ethical concerns (Tonkens, 2012). We leave aside the question of whether we need or should develop robots even partially morally independent of humans (Hew, 2014), or whether such an attempt would be permissible from the point of view of a given ethical framework (Tonkens, 2012), or whether a machine ethical framework should in fact fully reflect any human one. Although some researchers point out that the key aspect of virtue ethics lies in the deliberation of virtues (for example (MacIntyre, 2007)), we argue that, for the purpose of implementing ethical decision-making in autonomous devices, we have to focus on this process assuming that virtues are given (declared in advance).

Secondly, we have to explain why we choose to use a knowledge-based mechanism instead of the largely popular machine learning-based ones. The main reason for our choice is that ML-based systems function in the so-called black-box style, and they do not allow for explaining their knowledge and decisions. Although there are some approaches to model ethical behavior in machine learning-based AI systems (Abel, MacGlashan, & Littman, 2016), it is worth noticing that the lack of possibility of explaining decisions and the uncertainty of reasons for the decisions made by the ML-based devices is one of the key disadvantages of the ML-based autonomous devices. Although some of the researchers argue that machines can be taught ethical or legal behaviour via machine learning techniques (Webb et al., 2020), others (like (Steging, Renooij, & Verheij, 2021)) prove that good results can be made on the basis of wrong reasons.

On the basis of the above, we assumed that the part of the system responsible for ethical decision-making should be constructed on the basis of the knowledge-driven paradigm rather than the data-driven one. Such an approach allows us to address the value-alignment problem by direct introduction of the set of values which have to be satisfied for a decision to be moral. This does not exclude the possibility of using ML-based mechanisms in other elements of autonomous devices like decision options extraction, prediction of their results, and their evaluation in the light of relevant values. Our model is then consistent with Wallach and Allen's viewpoint (Allen & Wallach, 2012), as they point out that virtue ethics in autonomous devices should consist in a hybrid approach, joining together bottom-up and top-down techniques. However, the technical details of the ML-based modules are beyond the scope of this paper.

It should be noted at this point that although we operate within a virtue ethics framework for our model, some of its elements could be seen as transgressing this approach. Attempts to apply ethical frameworks to devices have not infrequently resulted with hybrid models (Tolmeijer et al., 2021); admittedly, our approach does incorporate some elements of other approaches. It could be viewed as partially consequentialist in the sense that actions' results are also considered in the assessment of decision options.

It is also important to set boundaries of our work. One of the most important assumptions is that, our model is focused on the ethical decision making, but not the ethical deliberations. In other words, our model is constructed on the basis of virtues declared in advance and does not introduce any mechanism of automatic generation of virtues or learning what is virtue. Such an assumption is rooted in the opinion that an autonomous device should follow our (human) ethics and morality, rather than work out its own "morality".

Finally, although the concept of value is essential for our model, at this point we remain value agnostic, in the sense that we do not engage into a discussion on which values should be used in a given type of autonomous device or what the source of values or relations between them ought to be.

GVR model

Our model of virtue ethics is constructed with the use of the model of value-based teleological reasoning from (Zurek, 2017), further referred to as the GVR (Goals, Values, and Reasoning) model. Below we present its summarized discussion. Firstly, the naming convention will be presented:

- By upper case letters we denote sets;
- By lower case letters we denote propositions;
- Subscripts denote names of propositions;
- Superscripts denote names of sets;
- Greek letters denote functions;
- Other symbols will be defined later (except trivial logical and set-theory ones).

Definition 1 (State of affairs)**.** Let $S = \{s_0, s_1, s_2, ...\}$ be a finite, non-empty set of propositions. Each proposition represents one state of affairs. Let γ be a function which returns 1 if a given state of affairs is true and 0 if not. One and only one element from set S can be true: if $(\gamma(s_y) = 1)$, then $\forall_{s_x \in S: s_x \neq s_y}(\gamma(s_x) = 0)$. s_0 is the initial state of affairs.

We assume that every single state of affairs ($s_x \in S$) represents one complete factual situation.

Definition 2 (Actions)**.** As an action we understand an activity which carries a transition from a certain state of affairs to another state of affairs. Actions will be represented by propositions from set $A = \{a_1, a_2, \ldots a_k\}$.

Note that a particular action cannot be performed in every state of affairs. The set of all possible actions in all possible states of affairs we denote as AS ($AS \subseteq A \times S$). Set AS is a set of pairs $AS = \{as_{i,j}, as_{k,w}, ...\}$ in which $as_{i,j} = \langle a_i, s_j \rangle, as_{k,w} = \langle a_k, s_w \rangle$ (the first subscript denotes the name of an action, the second subscript denotes the name of a situation). Each pair represents that a given action (for example, a_i) can be performed in a given state of affairs (for example, s_j). By AS^j (where $AS^j \subset AS$) we denote a set of actions possible to perform in a state of affairs s_j.

Function $\delta : AS \to S$ returns the result of performing an action a_i in a state of affairs s_j. By $\delta(as_{ij}) = s_y$ where $as_{ij} = \langle a_i, s_j \rangle$ we denote that the result of performing an action a_i in a state of affairs s_j is s_y.

Definition 3 (Transition process)**.** Let $\varepsilon : AS \times S \to S$ be a partial function which represents performing an action a in a state of affairs s. If $\delta(as_{i,j}) = s_y$ and $\gamma(s_j) = 1$ (the result of performing an action a_i in a state of affairs s_j is s_y), then performing $\varepsilon(as_{i,j})$ causes changing $\gamma(s_j) = 0$ and $\gamma(s_y) = 1$.

Definition 4 (Situation)**.** A situation is either a particular state of affairs or a result of a particular action performed in a given state of affairs. A set of situations X is a union of sets of states of affairs and results of actions: $X = \{S \cup AS\}$. By $x_n \in X$ we denote an element from set X. By X^j we denote a set of situations available from a state of affairs s_j: $S^j = \{s_j \cup AS^j\}$ (we introduce the concept of situation because state of affairs and actions can become decision options and it is easier to use symbol X instead of $\{S \cup AS\}$)

Definition 5 (Values)**.** We have to separate the two meanings of the word value: a value may be understood as a concept or as a process.

1. Value as an abstract concept which allows for the estimation of a particular action or a state of affairs and influences one's behaviour. V is a set of values: $V = \{v_1, v_2, \ldots v_n\}$

2. Value as a process of estimation of the level of extent to which a particular situation (state of affairs and / or action) x promotes a value v_i. By $v_i(x)$ we denote the extent to which x promotes a value v_i[1]. By $V(X)$ we denote the set of all valuations of all situations.

It is important to emphasize that values can be promoted (satisfied) to a certain degree by a particular state of affairs or action: $v(as_{i,j})$ represents the degree to which a value v is promoted by a state of affairs which is the result of performing an action a_i in a state of affairs s_j.

Although, similar to (Zurek, 2017), we are not going to impose here any particular way of representing the levels of promotion of values, we are not excluding to represent them as numbers. For example, in (Zurek & Mokkas, 2021) they were expressed by numbers from range $\langle 0; 1)$[2].

By $V^i(X)$ we denote the set of all possible extents to which a value v_i from set V may be promoted by any possible situation $x \in X$.

A partial order $O_i = (\geq; V^i(X))$ represents the relation between extents to which values are promoted: $v_i(x_n) \geq v_i(x_m)$ means that $x_n \in X$ promotes a value v_i to a no less extent that $x_m \in X$. If $v_i(x_n) \geq v_i(x_m)$ and $v_i(x_m) \geq v_i(x_n)$, then extents to which a situation x_n and x_m promotes a value v_i are equal ($v_i(x_n) = v_i(x_m)$). If $v_i(x_n) \geq v_i(x_m)$ and $v_i(x_n) \neq v_i(x_m)$, then $v_i(x_n) > v_i(x_m)$.

In real-life reasoning people do not rely only on a comparison of the levels of promotion of one value; usually, they compare the levels of promotion of various values. Theoretically speaking, they are incompatible, but practically, people compare not only the levels of promotion of various values, but also the levels of promotion of various sets of values.

Definition 6 (Sets of values)**.** By $V^Z \subset V$ we denote a subset (named Z) of a set of values V which consists of values: $v_i, v_j, ... \in V^Z$.

By $V^{x_i} \subset V$ we will denote a set of values promoted by a situation X_i.

Definition 7 (The level of promotion of a set of values)**.** By $V^Z(x_n)$ we denote a set of estimations of the levels of promotion of values constituting set V^Z by a situation $x_n \in X$. If $V^Z = \{v_z, v_t\}$, then $V^Z(x_n) = \{v_z(x_n), v_t(x_n)\}$.

By $V^{x_i}(x_i)$ (when the upper script contains the name of a situation) we denote a set of estimations of the levels of promotion of all values promoted by a situation $x_i \in X$.

Definition 8 (Value-extent preference)**.** A partial order $OR = (\triangleright; 2^{V(X)})$ represents a preference relation between various values and various sets of situations: $V^Z(x_n) \triangleright V^Y(x_m)$ means that the extent to which values from set V^Z are promoted by a situation x_n is preferred to the extent to which values from set V^Y are promoted by a situation x_m.

How to determine whether the extents to which all values promoted by one situation are preferred to the extents to which all values are promoted by another situation? This is not a trivial issue, because we have to balance between different levels of promotion of different values. This problem has been extensively discussed in (Zurek, 2017) and (Zurek & Mokkas, 2021), where two different approaches to this problem were presented. Although we believe that this topic still requires further development, it is outside the scope of our paper.

In our model, inference is based on the defeasible inference rules (by defeasibility we understand the possibility of defeating the conclusion obtained with the use of such rule by other, stronger, rule) represented by argumentation schemes(Bex,

[1] Note that a given situation is evaluated as whole: not only consequence of an action but also action itself, circumstances, etc.

[2] This means that the level of value promotion can be equal to 0 (no promotion at all), but can never reach 1 (a given value cannot be promoted to the maximal level).

Prakken, Reed, & Walton, 2004). Framework for reasoning and inference mechanism for GVR model was introduced in (Zurek & Mokkas, 2021). Inference rule has an antecedent part and a consequence part, separated by a double bar which stands for the sign of defeasible inference[3]. For the sake of brevity we present only one rule:

AS2 **Generalized practical reasoning:**[4] If in circumstances s_m performing an action a_t is preferred to remaining in s_m, $as_{t,m}$ is also preferred to any situation available from states_m, and $as_{t,m} \in AS$, then an action a_t should be performed:

$$\begin{array}{c} \exists_{s_m \in S} \exists_{as_{t,m} \in AS} \forall_{as_{k,m}} : \\ VO^{as_{t,m}}(as_{t,m}) \rhd VO^{s_m}(s_m) \land \\ VO^{as_{t,m}}(as_{t,m}) \rhd VO^{as_{k,m}}(as_{k,m}) \\ \hline\hline \varepsilon(as_{t,m}) \end{array}$$

Virtue in a computational model

The essential concept of the analyzed ethical theory is virtue. A virtue (arête) can be understood as aptitude / disposition / an excellent trait of character, which makes a person good. One of the key problems in the formal modeling of virtue ethics is to represent virtue in a computational way. In our model, we introduce values as an intermediate concept connecting virtues with particular actions, decisions, and states of affairs. Value is an abstract (trans-situational) concept which allows for the estimation of a particular action or a state of affairs and influences one's behavior. Consequently, on the basis of such a definition, we assume that particular values can be satisfied to a certain degree. Values can be seen as abstractions of particular situations and they can be promoted (or satisfied) to a particular extent by those situations (state of affairs, actions, etc.). Such an understanding of values allows us to treat them as a motivational aspect of one's behavior: the agent makes a decision *in order* to promote a particular value or set of values. How do we understand a virtue in this context? Since virtue means aptitude, disposition, or an excellent trait of character, we can represent it as a set of thresholds of a set of values relevant to the virtue. The agent's decision will be moral if it will be coherent with a virtue, i.e. the levels of satisfaction of values by this decision will be above the thresholds following from a given virtue.

Such a representation of virtue ethics expands the explanatory power of an autonomous agent by taking into consideration the motivational aspect of decisions and relating it to the agent's moral attitude. It also allows for taking a higher level of control over the devices. Such control focuses on revising the motivations of decisions rather than analysing them from the point of view of deontic modalities of actions or a state of affairs they bring about.

GVR and virtue ethics

Eudaimonist virtue ethics is rooted in ancient Greek philosophy, especially in the works of Aristotle. Although it is probably the oldest of the major ethical paradigms, it is the least implemented one in the light of possible application to autonomous devices. The human-centered character of this theory and the ambiguous character of the definition of virtue are the most important reasons for the lack of successful models of virtue-based ethical decision-making.

In order to create a model of virtue-based ethical decision making system, it is necessary to introduce a clear definition of what we understand as virtue and what does it mean that a particular decision or act exemplifies a virtue.

We realize that we attempt to formalize a very abstract and obscure concept, but we are convinced that the formalization of the concept of virtue is necessary for the sake of our goal. The main assumptions on the basis of which we can introduce and formalize the concept of virtue are: (1) a virtue represents a human's desired moral attitude, (2) virtue should be related to values or, in other words, values represents virtues, (3) a given decision option can satisfy or not a given virtue, (4) a decision is ethical if it satisfies a virtue or virtues. One can find the above assumptions as constraining the initial meaning of this term, but we believe that such simplifications are inevitable in successful modeling of complex real life concepts.

On the basis of the above assumptions, we can assume that a virtue can be represented by a set of minimal extents to which the decision should promote a particular set of values.

Definition 9 (Virtue)**.** Let virtue will be represented by the minimal extents to which a particular situation should promote a given set of values.

In a formal way:

- $VIRTUES = \{vrt_1, vrt_2, ...\}$ - denotes a set of virtues.
- By $v_n min(vrt)$ we denote the minimal extent (threshold) to which the promotion of a value v_n satisfies a virtue vrt.
- By $v_n(x_1) \geq v_n min(vrt)$ we denote that a virtue vrt is satisfied by a situation x_1 with respect to a value v_n.
- By $v_n \in vrt$ we denote that the minimal extent of a given value v_n is declared in a virtue vrt: $v_n \in vrt \leftrightarrow \exists v_n min(vrt)$.

In other words, virtue can be seen as a kind of abstract goal, but the goal which does not represent the agent's personal needs or desires, but the ideal moral attitude.

Having a model of virtue, it is possible to define when a particular decision option (situation) will satisfy a given virtue:

Definition 10 (Satisfaction of a virtue)**.** Assuming a situation $x_i \in X$ and virtue $vrt_j \in VIRTUES$, situation x_i will satisfy virtue vrt_j: $Vsat(x_i, vrt_j)$ iff
$\forall_{v_n \in vrt} : v_n(x_i) \geq v_n min(vrt_j)$

[3] To preserve the cohesion of designations, the names of argumentation schemes will be the same as used in (Zurek, 2017).

[4] This is a modified version of the AS2 from (Zurek & Mokkas, 2021).

The definition of satisfaction of a virtue relates to only one virtue. Let $VSAT(x_i)$ denote that a particular decision option satisfies all virtues:

Definition 11 (Satisfaction of all virtues)**.** We say that a particular situation $x_i \in X$ satisfies all virtues iff:
$VSAT(x_i)$ iff $\forall_{vrt_i \in VRT} Vsat(x_i, vrt_i)$

The above definitions allow us to evaluate a particular decision in the light of virtues, hence we can say that a particular decision exemplifies a given virtue if this virtue is satisfied by the situation. Additionally, we can say that a particular situation is moral if it satisfies all virtues.
The possibility of evaluation of a particular situation in the light of virtues allows us to update the existing inference rules in order to model moral reasoning:

AS2' **Generalized moral practical reasoning:** If in circumstances s_m performing an action a_t is preferred to remaining in s_m, a_t is preferred to other actions available from S_M, $as_{t,m} \in AS$, and $as_{t,m}$ satisfies all virtues, then action a_t should be performed:

$$\frac{\begin{array}{c}\exists_{s_m \in S} \exists_{as_{t,m} \in AS} \forall_{as_{k,m}} s.t.VSAT(as_{k,m}) : \\ VO^{as_{t,m}}(as_{t,m}) \rhd VO^{s_m}(s_m) \\ VO^{as_{t,m}}(as_{t,m}) \rhd VO^{as_{k,m}}(as_{k,m}) \land \\ VSAT(as_{t,m}) \end{array}}{\varepsilon(as_{t,m})}$$

From a general standpoint, one can notice that in our model, decisions which do not satisfy the requirements of virtues are a kind of "no-go zones." This, of course, raises a question of what if there are no available decisions which fulfill the virtues' requirements? Such a question relates to the assumed policy: what to do if there are only bad decisions available? The answer to this question is outside the scope of this paper, because it is not related strictly to virtue ethics (this question may appear in the models of other ethical theories), but rather to the adapted decision making policies and procedures. Leaving this discussion aside, it is possible to add a kind of a lesser evil argument to our model for practical reasons (e.g. (Zurek, Araszkiewicz, & Stachura-Zurek, 2022)).

Related work

Despite a significant number of formal models of deontic and consequentialist ethics for autonomous devices (see (Tolmeijer et al., 2021) for a survey) proposed so far, few such models exist in the domain of virtue ethics. Below we briefly present how the approach presented in this paper relates to the existing discussion of virtue ethics in autonomous devices.

The problem of ethical decision making in military systems was discussed in (Arkin, 2007), where the author presents a very general model of making military decisions following moral and legal principles. The core idea of the model is the filtering of illegal or non-moral behaviors. The author focuses on a deontic approach, with minor elements of consequentialist ethics. The author mixes legal rules with ethical ones, treating ethical principles as an element of legal system; in consequence, he rejects virtue ethics as non-suitable for the representation of codified ethics. Such an approach seems an oversimplification of the problem of ethical decision making, as not every ethical problem can be easily modelled with the use of a deontic approach.

The author of (Coleman, 2001) presents a general discussion of the problem of definition of virtue in the context of AI-based agents, thus defining virtues: *Virtues are those qualities or characteristics which enable and foster an agent's pursuit and achievements of its end.* While such a definition emphasizes the teleological character of virtues, resembling our approach in this respect (in our model, satisfying virtues is one of the goals of agent), in this paper we further explore the understanding of what "quality" or "characteristics" means, by saying that such a quality is a minimal acceptable level of a satisfaction of a value. Our approach is then on the one hand more specific in representing the nature of virtue, but on the other hand it does not take into consideration non-functional, technical properties of the device (reactiveness, freedom from bias, etc.) Coleman also distinguishes some properties of virtue ethics and translates them into a domain of computational agents (Coleman, 2001): (1) Virtuous agents are ideal or excellent agents, (2) Virtues are properties that make such agents ideal, (3) Virtues are the means to some end, (4) Virtues are often thought of as being unified (i.e., you either have all or none). In our work, the excellence is made possible by fulfilling the virtues, i.e. satisfying values to the required level (which is fulfilment of a goal), such virtuous agents are excellent (ideal) because they fulfill requirements.

The problem of implementation of virtue ethics in machines has been discussed in (Berberich & Diepold, 2018). The authors argue that reinforcement learning mechanisms are perfect tools for implementing virtue ethics in autonomous devices, because they are, similar to humans, trained to behave virtuously. The representation of the reward function can, in their opinion, represent a virtue, but the authors do not discuss how in particular such a function would be represented. While agree that such an approach is similar to how humans learn to be virtuous, the key (and unsolved) difficulty lies in the representation of virtue in the reward function. Moreover, even if such a function can be created, the authors cannot, due to the specific nature of deep reinforcement learning properties (especially the lack of transparency), be sure that the machine will "learn" the right properties of being virtuous. This is connected with the so-called value alignment problem. Although the authors of (Berberich & Diepold, 2018) point out that their model solves this problem, they do not provide any convincing justification for such a claim.

An interesting discussion of the possibilities of implementation of virtue ethics in machines has been presented in (Stenseke, 2023). The author discusses the concepts of virtue, eudaimonia, and phronesis in the light of the possibility of representing them in artificial devices. He argues that a connectionist approach (neural networks) is the most suitable tool for representing virtue ethics in machines. Although our model is constructed on the basis of a different paradigm (we are using a knowledge-based approach instead of a data-driven one), there are some interesting relations in the understanding of particular components of virtue ethics in both models. The concept of *virtue* in (Stenseke, 2023) is understood as a kind dimension on the basis of which a decision (action) is evaluated, while virtue in our work has a more abstract character, representing the extents to which a set of values should be promoted (technically, virtue in (Stenseke, 2023) is value in our model). The concept of *eudaimonia* in (Stenseke, 2023) is in fact the reward function: the function which should be maximized. Our understanding of *eudaimonia* is different: following Aristotle, we understand *eudaimonia* as particular state of affairs which relates to the human ultimate state of happiness. In other words, in our model *eudaimonia* is a kind of a goal state, while in (Stenseke, 2023) *eudaimonia* is a direction only. In addition, the concept of *phronesis* in both models is not treated interchangeably: in our model, *phronesis* is represented by the knowledge of the reasoning mechanism, while (Stenseke, 2023) divides it into 2 parts: (1) learning what action leads to good and (2) learning what is good in itself; the assumption is also that both parts should be learned. Although we can agree that the knowledge about which action is good can be obtained by learning, we argue that the knowledge of what is good should be directly introduced into the machine. This element of the model is informed by one the basic assumptions of our work: the wish that machines follow *our* ethics and evaluations of what is good (particularly relevant in deep learning mechanisms which face the so-called value-alignment problem), clearly highlighting how these two approaches to machine ethics implementation diverge. While (Stenseke, 2023) tries to reflect or simulate the human-like virtue ethics in a machine context as accurately as possible, this work proposes a model of the virtue ethics-based decision-making mechanism which uses some concepts of virtue ethics to make moral decisions.

The authors of (Hegde, Agarwal, & Rao, 2020) analyze the behaviour of agents equipped with elements of utilitarian and virtue ethics in the Continuous Prisoner's Dilemma. Every agent in the experiment has two parameters: resource and reputation. The overall evaluation by the agent is made on the basis of both utilitarian and virtue ethics premises. (Hegde et al., 2020) shares some elements similar to our model (the utilisation of a threshold in the representation of the virtue ethics), but the overall presentation is much more simplistic. Moreover, it is not quite clear how the authors understand a virtue: if a virtue is connected with the action (as claimed by the authors), the paper could arguably exemplify a deontological approach.

One of the most important models of consequentialist and virtue ethics is discussed in (Bench-Capon, 2020). The model is constructed on the formal basis of the AATS+V model (presented in (Chorley & Bench-Capon, 2005), (Atkinson & Bench-Capon, 2007), (Atkinson & Bench-Capon, 2014), and other). Bench-Capon's approach (Bench-Capon, 2020) revolves around the assumption that virtue is an ordering between values and offers an understanding of the concept of virtue different from the one adapted in this paper: while in our model, virtue is a set of thresholds of the levels of values' promotion, in Bench-Capon's approach virtue is a hierarchy between values. In our opinion, representation of virtue as a set of thresholds is much more useful, because it prevents the autonomous device from decisions in which the strong promotion of one value too strongly decreases the level of promotion of the other value.

Concluding discussion

The model presented in this paper has been created within the eudaimonist ethical framework, where the fulfilment of the human potential, full development of oneself as a person - eudaimonia - can be achieved as a result of virtuous life. The focal point of this approach is the concept of virtue. In our model, by a virtue we understand the minimal acceptable levels of promotion of a set of values. This understanding of a virtue is naturally a simplification of this concept, but it allows for representing the key elements of virtue ethics in autonomous devices.

We assumed that virtues are the basis for decisions and eudaimonia is achievable. This view on virtues and eudaimonia implies the binary character of these concepts: virtue can be fulfilled (or not) and eudaimonia can be achieved (or not). Considering that a virtuous life should lead to eudaimonia, we can assume that if a device makes decisions which fulfill the virtues, it achieves eudaimonia (note that by eudaimonia we do not understand "eudaimonia" of a machine, but eudaimonia of humans). In other words, if a device makes a decision which promotes values to the levels above the thresholds established by a virtue, then it is moral, i.e., brings about eudaimonia.

Another key concept of virtue ethics is phronesis, understood as practical wisdom or – more suitably in this case – the capacity to reason about virtues. In our model this concept is represented by the entire reasoning mechanism, on the basis of which a device can infer which decision leads to eudaimonia, i.e., which decision satisfies all virtues.

Moreover, thanks to our model, we can point out what kind of technological developments are needed to promote the responsible use of AI from a socio-technical perspective. In our opinion, the key challenge lies not in the problem of ethical reasoning itself, but rather in the evaluation of the available decisions. From a technical point of view, such a task can be understood as a kind classification or regression task, the leading question for future research being whether the creation of such a regression mechanism is feasible at all.

References

Abel, D., MacGlashan, J., & Littman, M. L. (2016). Reinforcement learning as a framework for ethical decision making. In B. Bonet et al. (Eds.), *Aaai workshop: Ai, ethics, and society* (Vol. WS-16-02). USA: AAAI Press. (978-1-57735-759-9)

Allen, C., & Wallach, W. (2012). Moral machines: Contradiction in terms or abdication of human responsibility? In *Robot ethics: The ethical and social implications of robotics* (p. 55-68). Cambridge, USA: MIT Press.

Arkin, R. (2007). *Governing lethal behavior: Embedding ethics in a hybrid deliberative/reactive robot architecture* (Tech. Rep. No. GIT-GVU-07-11). USA: College of Computing Georgia Institute of Technology.

Atkinson, K., & Bench-Capon, T. (2007). Practical reasoning as presumptive argumentation using action based alternating transition systems. *Artificial Intelligence*, *171*(10-15), 855 - 874.

Atkinson, K., & Bench-Capon, T. (2014). States, goals and values: Revisiting practical reasoning. In *Proceedings of 11th intl. workshop on argumentation in multi-agent systems.*

Bench-Capon, T. (2020). Ethical approaches and autonomous systems. *Artificial Intelligence*, *281*, 103239. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0004370219300621` doi: https://doi.org/10.1016/j.artint.2020.103239

Berberich, N., & Diepold, K. (2018, 06). The virtuous machine - old ethics for new technology?

Bex, F., Prakken, H., Reed, C., & Walton, D. (2004). Towards a formal account of reasoning about evidence: Argumentation schemes and generalisations. *Artificial Intelligence and Law*(11), 125 – 165.

Burbules, N. C. (2019). Thoughts on phronesis. *Ethics and Education*, *14*(2), 126-137. Retrieved from `https://doi.org/10.1080/17449642.2019.1587689` doi: 10.1080/17449642.2019.1587689

Chorley, A., & Bench-Capon, T. (2005). An empirical investigation of reasoning with legal cases through theory construction and application. *Artificial Intelligence and Law*, *13*(3-4), 323–371. doi: 10.1007/s10506-006-9016-y

Coleman, K. G. (2001). Android arete: Toward a virtue ethic for computational agents. *Ethics Inf Technol*(3), 247-265.

Hegde, A., Agarwal, V., & Rao, S. (2020, 7). Ethics, prosperity, and society: Moral evaluation using virtue ethics and utilitarianism. In C. Bessiere (Ed.), *Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20* (pp. 167–174). USA: International Joint Conferences on Artificial Intelligence Organization. Retrieved from `https://doi.org/10.24963/ijcai.2020/24` doi: 10.24963/ijcai.2020/24

Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics Inf Technol*(16), 197-206.

Hursthouse, R., & Pettigrove, G. (2018). Virtue Ethics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2018 ed.). USA: Metaphysics Research Lab, Stanford University. `https://plato.stanford.edu/archives/win2018/entries/ethics-virtue/`.

MacIntyre, A. (2007). *After virtue: A study in moral theory.* University of Notre Dame Press. Retrieved from `https://books.google.de/books?id=7bLuAAAAMAAJ`

Russell, S. (2019). *Human compatible: Artificial intelligence and the problem of control.* UK: Penguin Publishing Group. Retrieved from `https://books.google.pl/books?id=M1eFDwAAQBAJ`

Saja, K. (2015). *Etyka normatywna.* Krakow: Universitas.

Sim, M. (2007). *Remastering morals with aristotle and confucius.* UK: Cambridge University Press.

Steging, C., Renooij, S., & Verheij, B. (2021). Rationale discovery and explainable AI. In S. Erich (Ed.), *Legal knowledge and information systems - JURIX 2021: The thirty-fourth annual conference, vilnius, lithuania, 8-10 december 2021* (Vol. 346, pp. 225–234). Netherlands: IOS Press. doi: 10.3233/FAIA210341

Stenseke, J. (2023). Artificial virtuous agents: from theory to machine implementation. *AI & Soc*, *38*, 1301–1320. doi: https://doi.org/10.1007/s00146-021-01325-7

Swanton, C. (2015). *The virtue ethics of hume and nietzsche.* USA: Wiley-Blackwell.

Tolmeijer, S., Kneer, M., Sarasua, C., Christen, M., & Bernstein, A. (2021). Implementations in machine ethics: A survey. *ACM Comput. Surv.*, *53*(6). Retrieved from `https://doi.org/10.1145/3419633` doi: 10.1145/3419633

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

Webb, N., Smith, D., Ludwick, C., Victor, T., Hommes, Q., Favaro, F., ... Daniel, T. (2020). *Waymo's safety methodologies and safety readiness determinations.*

Yu, J. (2007). *The ethics of confucius and aristotle: Mirrors of virtue.* USA: Routledge.

Zurek, T. (2017). Goals, values, and reasoning. *Expert Systems with Applications*, *71*, 442 - 456. doi: http://dx.doi.org/10.1016/j.eswa.2016.11.008

Zurek, T., Araszkiewicz, M., & Stachura-Zurek, D. (2022). Reasoning with principles. *Expert Systems with Applications*, *210*, 118496. Retrieved from `https://www.sciencedirect.com/science/article/pii/S0957417422015792` doi: https://doi.org/10.1016/j.eswa.2022.118496

Zurek, T., & Mokkas, M. (2021). Value-based reasoning in autonomous agents. *International Journal of Computational Intelligence Systems*, *14*, 896-921. doi: https://doi.org/10.2991/ijcis.d.210203.001

Zurek, T., & Stachura-Zurek, D. (2021). Applying ethics to autonomous agents. In J. Bylina (Ed.), *Selected topics in applied computer science* (p. 199-222). Lublin, Poland: Wydawnictwo UMCS. (`https://wydawnictwo.umcs.eu/js/elfinder/files/Ebook/Selected%20Topics%20in%20Applied%20`

Computer%20Science%20%28vol.%20I%29.pdf)