Under review manuscript No.
(will be inserted by the editor)

Value Engineering for Autonomous Agents – Position Paper

blinded

Received: date / Accepted: date

Abstract Ethics in Artificial Intelligence is a wide-ranging field which encompasses many open questions regarding the moral, legal and technical issues that arise with the use and design of ethically-compliant autonomous agents. Under this umbrella, the computational ethics area is concerned with the formulation and codification of ethical principles into software components. In this position paper, we take a look at a particular problem in computational ethics: value engineering in autonomous agents. This work aims at building the philosophical foundations that a future model of value engineering should be based on. The main points of our proposal are: (1) values are introduced into agents as goals that ground the meaning of those values; (2) norms are the means to steer an agent society towards beneficial outcomes, and hence should be used to promote values; and (3) autonomous agents can negotiate over norms to align them with their values, in an exercise of value aggregation. Finally, we argue that our proposal does not endow software agents with moral agency, as there is always a human team responsible for deciding which values should be encoded and the meaning they take, i.e. the form of the grounding goals. We believe that this position paper accounts for a solid philosophical foundation for a future formal model of values in autonomous agents, and provides the starting points for work in that direction.

Keywords Computational ethics · Value engineering · Value alignment · Normative systems · Multiagent systems · Philosophical foundations of values in AI

1 Introduction

In recent years, many works have been developed under the banner of ethics in artificial intelligence (AI). These range from philosophical investigations on the moral agency of autonomous entities [8,5], legal issues related to their autonomy and accountability [3], ethical concerns with regards to the behaviours that more powerful technologies enable [1] and the technical realisation of ethically-compliant autonomous agents [2].

Address(es) of author(s) should be given

Within this large and diverse area, the field of computational ethics has been recently outlined to include efforts on turning ethics into computable entities, and the study of its complexity and tractability [7, 14].

Computational ethics deals with the formulation and implementation of models that codify abstract moral principles and theories into computer programs. At the very least, building and embedding ethics in a systematic manner into software systems allows for the automatic verification of the system's compliance with the formulated moral principles in a rigorous way. If such models are introduced into agents endowed with autonomy and interactivity, they could allow for the agents to adapt to the value requirements imposed, and minimise the need for human intervention whenever the ethical requirements change, avoiding lengthy discussions among stakeholders and redesign operations.

Despite its prominent practical aspect, computational ethics is not detached from the philosophical discussions on values and morality that underlie the formal models and programming approaches that will be realised in practice. Furthermore, making such assumptions explicit early on in the development process ensures the robustness of the formulation, facilitates discussions on its possible weaknesses and helps frame the research within the larger landscape.

In this work, we study a particular problem in computational ethics, which is the definition of the philosophical foundations for a formal model for value engineering in autonomous agents. We do not intend to go into the details of the mathematical formulation here, but to present the philosophical and psychological foundations of the theory of moral values we adopt, and that could underpin the mathematical basis of the AI community's work on operational ethics. We start off at Schwartz's Theory of Basic Human Values to establish the nature of values as formal concepts, how they relate to the world that agents populate and the function they serve. Then, we argue for the potential of prescriptive norms as the main value-promoting mechanism, and provide an outline of an agent architecture that would allow agents to actively analyse, adopt and promote norms based on their success at upholding the values they most esteem.

This position paper is organised as follows. In Section 2 we review the main points on value theory that we intend to export into our model of computational ethics. Then, we explain how such an adaptation of values from the social science domain to the technological domain should happen. In Section 3, we make the case extensively for the central positions that norms ought to have. In Section 4 we make the point that, despite all the ethically motivated capabilities we intend to provide agents with, we still do not consider them to have moral agency. Finally, in Section 5 we outline the main features of our proposal and point to the future work that should be built upon it.

2 Values as formal objects

2.1 Values in the social sciences

Currently, Schwartz's Theory of Basic Human Values is one of the most widely accepted frameworks on moral values in psychology and sociology. The main success of this theory has been in delineating a set of broad value types that provide motivation for individuals and social groups across all major cultures. This set has been found to be comprehensive, meaning that any explanation provided by an individual as justification for her behaviour can be related to one of the identified types. Additionally, the meaning of these values and the relationships of conflict and compatibility among them were also found to be largely consistent both across cultural groups and within the members of the same society.

The main impact area of Schwartz's theory has been in intercultural studies, however, for the matter at hand (a model of computation ethics for autonomous agents) our interest is not on the achievements of the theory but on the conceptualisation of values that it works with. Specially because its definition of the concept of *values* and their function is quite standard among the social science literature. We draw inspiration from it since we want our modelling to be consistent with established research in those areas, and we expect to export it into the realm of autonomous agents to the extent possible. According to [13, p. 4]:

> Values (1) are concepts or beliefs, (2) pertain to desirable end states or behaviours, (3) transcend specific situations, (4) guide selection or evaluation of behaviour and events, and (5) are ordered by relative importance.

The theory conceives values as very general, abstract guiding principles that individuals and groups can appeal to when faced with ethically controversial situations. It also proposes that the content of every value is realised into a motivational goal:

> the primary content aspect of a value is the type of goal or motivational concern that it expresses. (...) values represent, in the form of conscious goals, three universal requirements of human existence to which all individuals and societies must be responsive: needs of individuals as biological organisms, requisites of coordinated social interaction, and survival and welfare needs of groups.

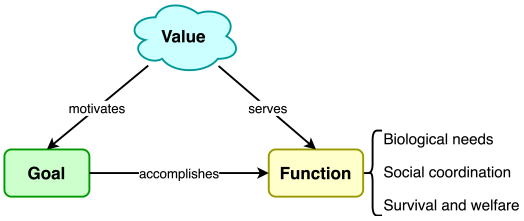

Fig. 1 The three main components of value theory and the relationships among them. Note the cloud-shaped box around values emphasising their abstract nature.

Hence, values, when activated, take the form of conscious goals whose function is to allow an individual and her community at large to thrive. Schwartz's theory of values is precisely composed of these three concepts: (1) moral values as abstract principles; (2) explicit goals that values motivate in specific situations; and (3) ultimate functions that those goals seek to achieve. Figure 1 presents a diagram with the relationships between the three concepts.

Of the three concepts, the one that is typically the less acknowledged is the function that values serve. When asked to justify why you paid for your groceries at the supermarket, most people would answer something like "because it is the right thing to do". Very few would state something along the lines of "refusing to pay for my items could motivate everyone else to do the same, throwing society into a down-spiral and eventually landing in a state of anarchy that could threaten my very existence". It is much more handy (and accounts for much shorter explanations) to make a moral argument and appeal to a shared sense of duty. This is also consistent with a method of acquiring values by copying our peers, instead of rationally analysing whether abiding by certain values will be evolutionarily beneficial every time we face an ethically controversial situation.

The differences between coming up with an immediate reason for paying for groceries in terms of right and wrong, or making an elaborate speculation of the consequences of shoplifting can be stated in terms of Kahneman's two-system approach to cognition. According to [6], any mental process is performed by either the fast, intuitive and low-effort System 1, or by the slow, calculating and consuming System 2. When someone is asked to justify her actions, the permanently alert System 1 provides an immediate answer by appealing to her values, and rapidly makes a coherent association (in Kahneman's words) from behaviour to its value abstractions. It is only when pressing the individual on the ultimate reasons for performing an act that the lazy System 2 wakes up, and its conscious reasoning may be able to effortfully link any everyday action to our most primal needs.

2.2 From humanities to technology

In the context of software agents, concerns about pure evolutionary survival do not really apply. Therefore, we focus on the other two components to build our model: the abstract values and the goals they motivate. According to Schwartz's theory, values are transcendental guiding principles, yet they manifest themselves in the form of conscious (or rather *explicit*) goals when they are activated in a particular context. A formal model of values, then, should start by accommodating this assertion: abstract values get grounded into permanent goals that agents actively pursue. Those may be sought both through actions that directly affect their environment and by crafting and implementing norms that facilitate states where these goals are fulfilled to a large degree (more on that in Section 3).

Hence, values are not "directly" mapped into computational entities, but the motivational goals that ground their meaning are. This feature of our model is exemplary of an intentional view of AI, where the system is provided with a target end and the reasoning machinery to get there. The task of the software agent is to deduce the

best course of action towards that end, if one is available. It should be noted that the sought-after outcome may include information on intermediate states and actions, and so our proposal should not be confused with an "ends justify any means" position. But we would like to remark its intentional nature, as opposed to extensive techniques, where the autonomous agent would be fed with a large database of ethically-approved and disapproved behaviour instances, conveniently annotated for the values they represent. Instead of specifying goals as input, extensional methods provide data on the actions.

Without going into much detail, we perceive two main problems with extensional approaches to ethics in autonomous agents. The first one is related to the hidden nature of motivational goals that is implicated when learning by example. By working just with instances of ethical behaviour, the objective that any action is pursuing in the context where it is performed is very much implicit, and possibly not even known to the human designer. This can easily lead to problems related to the control of the system, which might have disastrous consequences in high-stakes situations. The second problem has its origin in the dynamic relationships that values have with one another. An action can be simultaneously compliant and/or disregarding of several values, and in varying degrees. The "moral annotation" of a behaviour instance becomes complicated when one acknowledges that values do not compose a set of mutually exclusive categories.

And so, in essence, we advocate for the introduction of permanent goals to achieve ethical behaviour in autonomous agents, both to retain a higher level of control (compared to extensional approaches) and facilitate the modelling of subtle interactions between values. A special feature of these goals is their permanent status. In contrast with goals in the classical AI tradition, the promotion of values is constantly sought after by the agents. They try to move closer towards them if the current state of affair is unsatisfactory or to perpetuate the desirable features of the present situation. Being in a state that is highly compliant towards a particular value is not enough, if subsequent transitions result in states where it is neglected. Because of their status as general guiding principles, the goals that values are grounded into do not, once achieved, disappear in the pursue of some other, more fundamentally desirable goal.

Another difference between the grounding goals motivated by values and traditional AI goals is their degree of satisfaction. Ordinarily in the planning literature, a goal is either fulfilled or it is not. When talking about moral values, the nuances of the concept do not really admit this dichotomy. For this reason, we propose that the goals that ground the meaning of values should not be evaluated to true/false, but rather to a degree of satisfaction over a continuous domain. In addition, we propose that this domain should be preferably bounded, with one end indicating perfect compliance towards the value in question, and the other reflecting complete neglect. Having a consistent grading scale across all values of interest will greatly enhance our ability to answer questions related to their compatibility or conflict.

So far, we have established that formally, values are grounded into permanent goals. Now we want to describe the cognitive tools that these values and their corresponding goals provide agents with. Values, through their grounding goals, operate in two distinct capacities. First, they work as evaluating devices that agents can resort to when they want to assess how desirable is the state of affairs, with respect to a value

or set of values. As mentioned, such an evaluation may include not just state variables that are instantaneously true, but also the history of events that has led to such state. Essentially, every agent has at its disposal a set of ranking criteria, one per value, that enable it to grade states by how successfully they promote every value (equivalently, how close they are to the goals that the values are grounded into).

Second, values can also be leveraged as guiding devices to help inform agents' decisions prior to executing some action. Again, such consideration may take into account how suitable the action is "by itself" (as is the habit of deontological ethics) and/or what is the foreseeable impact of the action on the state features, and whether those move closer or farther from the ideal situation (a position taken by utilitarian or consequentialist ethicists).

Despite the separation between the two uses of values, their distinction is somewhat superficial. The back-end computations for values both as guiding and as evaluation devices can be identical. The main difference resides in the context where they are employed. Values as guiding devices are mostly applied in situations where the system needs to determine the next action to be executed immediately. Differently, values as evaluation devices are intended for look-ahead planning tasks, where a complete sequence of actions is calculated, as well as for *what-if* analysis, whenever an agent needs to estimate in advance the effects of implementing a new norm (more in Section 3).

Finally, our model of computational ethics must account for the fact that the set of values observed by an agent, and by the community more broadly, do not play out in isolation. They are prioritised from the most esteemed to the least important. This feature is a direct import from the value hierarchy included in Schwartz's theory, and it becomes particularly relevant when conflicts between the goals of different values arise. An additional characteristic that we would like to introduce is the possibility for this hierarchy to be context-dependent. It is conceivable that circumstances might prompt agents to neglect some values that would otherwise be very important, and vice-versa.

3 The role of norms

In general, agents who interact in a shared environment will be subject to the norms regulating it. As formal entities, we understand norms in the prescriptive sense, as an assignment of a deontic modality to an action, alongside with the pre-conditions and post-conditions for such action. The subset of agents to whom a specific norm applies is among the most typical pre-conditions.

A unique characteristic of technical norms that is not shared with rules in the human domain is the possibility of regimentation. This means that, in some cases, compliance with some technical regulations can be perfectly accomplished, e.g. by eliminating the technical ability to perform some action. Non-regimented norms, however, are more interesting because they reflect our human reality more closely. Norms that seek to ban harmful behaviour but do not have the resources to enforce it perfectly usually rely on some form of punishment for detected offenders. As an

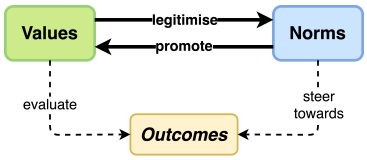

Fig. 2 Relationship between values and norms through outcomes.

example, we point to the temporary (and eventually, permanent) suspension of access to online communities to members who do not abide by the terms of use [4].

For our model, we do not make a commitment as to whether norms should be regimented or not. Our approach works with either. A fundamental characteristic of our proposal is that, regardless of their type, norms have a central role as the primary value-promoting mechanisms. Norms can, if carefully designed, facilitate the achievement of the goals that ground the meaning of values in the environment where the agents are operating. When implementing a new norm (or set of norms) leads to an outcome that is viewed as highly positive with respect to some value, we say that the norm *is aligned with respect to that value*. Hence the relationship between norms and values is consequential in nature. A norm is not moral in itself, it is so to the extent that the effects it brings about in the society agree with the members' values, represented in the form of goals.

Figure 2 represents the relationship between the two entities in schematic form. At the surface, values and norms form a feedback loop: values legitimise the enforced norms, and norms promote values when enforced. At a more fundamental level, the two are linked by the outcomes that norms steer the system towards and that are favourably evaluated with regards to values. By "outcomes" we include both the variables' values at an end state as well as the sequence of actions whose execution leads to it.

In general, the majoritarian approach to automated norms synthesis, whether ethical considerations are included or not, consists of an algorithm implemented outside of the system, which outputs a set of optimal and consistent regulations. There are recent works who share our view that norms should be automatically selected on the basis of the moral values they support, see e.g. [15]. Another line of work is Value-Sensitive Design (VSD), where the composition of morally adequate technical norms is hand-crafted by a human designer, and still made outside of the multiagent system.

We reject the dissociation between the multiagent system and the generation of the technical norms that regulate it of the previously mentioned approaches. This constitutes the most innovative feature of our proposal. To the best of our knowledge, only [9] has proposed an architecture for achieving the endogenous emergence of prescriptive norms through the participation of the agents. However, we are not aware of any follow-up on that work.

In our view, it should be the autonomous agents who attempt to align norms towards the values that the human designer has instilled in them. Prescriptive norms, then, with their explicit representation and syntax, should be handled by the agents populating the system, and evaluated by leveraging their understanding of values (i.e.

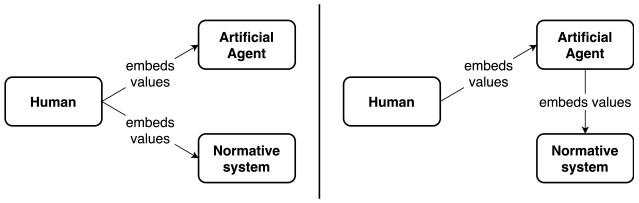

Fig. 3 Distinction in the embedding of values between Value-Sensitive Design (left) and our proposal (right).

the goals that values are grounded into) as evaluating devices. The designer, tasked with programming the agents, is not in charge of coding the technical norms directly. She is responsible, however, for including the necessary mechanisms that agents can resort to when crafting norms, figuring out the most probable outcomes that they lead to and ethically evaluating them.

Our approach constitutes a significant deviation both from the automated norm synthesis literature and the VSD world. With respect to the first case, we differ on the software component tasked with generating the norms (an outside algorithm vs the agents themselves). With respect to the second, we put a much stronger link between agents and norms. VSD assumes that the coding of the agent architecture and the technical norms are relatively independent tasks. In that framework, the human designer is directly responsible for embedding values into all the software components, in particular, into norms. In contrast, we propose that it should be the agents who craft, negotiate over and implement new norms on the basis of how well-aligned they perceive the candidate norms to be with respect to the values modelled in them. The difference between VSD and our approach is presented in schematic form in Figure 3.

A major concern in the field of AI is the possibility that building very sophisticated agents with a large degree of autonomy may backfire and end up hurting the humans whom they are supposed to serve. In our proposal, the main ambition is to leverage this autonomy and put it at the service of ethical behaviour. In comparison with the VSD approach, we expect the system to be more flexible. By collectively changing the norms in place, agents would have the ability to adapt to changes in the meaning of values (i.e. the goals that values get grounded into), their hierarchy and the introduction of new values. Note that the potential changes in the requirements are exogenous to the system, programmed by the designer, while the adaptation to those changes in the form of norms is endogenous.

The agents populating the system are instilled with values by a human who grounds their meaning as persistent goals. Every agent can be provided with its own, potentially different, version of grounding goals for the same value. Consequently, it is a human who has complete control over the meaning of values. To address possible differences among agents on the meaning of values, they need social skills to propose, communicate to others and eventually agree on the norms that are implemented on the entire system. We conceive the process by which agents collectively conform to a new prescription that was not part of the initial situation as a negotiation over the norm space.

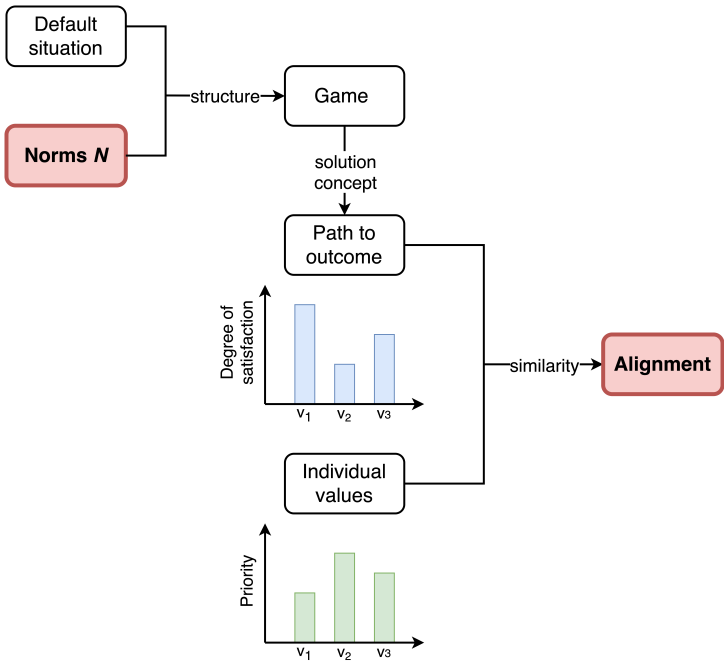

Fig. 4 Computational process for a single agent of the alignment of new candidate norms with respect to the values she regards the most.

When the negotiation process to add or modify the norms in place is triggered, the reasoning process that every agent follows is presented in Figure 4. Using its knowledge of the current system and its norms, plus the changes in norms being proposed, the agent is able to build a model of what would the interaction look like under the new set of norms in the form of a game. A solution concept (Nash equilibria, correlated equilibria, etc) is then applied to the game to yield a prediction of the path of play and the outcome that is expected to be reached. This forecast is then assessed by the individual in terms of the values that the agent regards, leveraging them as evaluating devices. Such an assessment would produce an assignment, for every value, of the degree of satisfaction the goal that it is grounded to. This is represented by the blue bar chart in Figure 4. The agent then compares this assessment to its value priority structure (stored in a data structure represented by the green bar chart). The similarity between the two, computed with an appropriate metric, yields the degree of alignment of the proposed norms with respect to the values held by that agent. The same computation is performed by every agent involved in the system.

The reasoning process just described allows an agent to position itself for or against a candidate norm once the negotiation process has started. The prediction is that the norms that come out of the bargain would correspond to a compromise between the value priorities held by the participating agents. We expect that agents with similar value structures would easily find the set of norms such that their alignment for all agents is above some moderate to high threshold. In the extreme case where all agents

share the same preferences over values and the same understanding of their meaning, the process should generate the optimal normative system with respect to all of them. Very heterogeneous societies could have a much harder time reaching an agreement over what regulations to adopt, or might not be able to agree on any at all.

An interesting perspective on the negotiation over norms is as a value-aggregating process. Agents come in to the bargain equipped with individual values, which might be very diverse both on their meaning and their priorities. From the negotiation process comes out a set of norms that modify the structure of the situation, and that are implemented on the system as a whole. The selection of norms takes into account the value preferences of the interested agents and produces a normative system that somehow merges all of them into a shared regulative body. This does not necessarily equate to having the resulting norms be optimal with respect to the average of the values of every separate agent. Whether norms are more responsive to a subset of agents over another will depend on how the negotiation process is set up, and how much power does each individual hold.

We illustrate the view of norm negotiation as a form of value aggregation with the following example. Consider a group of agents, each of them inculcated with the meaning of a set of values and their priorities. Let the agents negotiate and come to an agreement over the norms to be enforced. As mentioned already, we do not expect the resulting norms to be optimally aligned with respect to any of the agents. However, imagine a new special agent, who is not part of the initial group. We refer to it as the *socially equivalent agent*. The socially equivalent agent is equipped with an ethical structure (grounding goals plus value priorities) such that the norms agreed upon by everyone else *are* optimally aligned with respect to its value organisation. We recommend that the value structure of the socially equivalent agent should be referred to as the *social values* of the community. They do generally not reside in the data structures stored by any one individual, or subset of individuals, but emerge as a consequence of the interaction (i.e. the negotiation process) between individual values. Our future model of value engineering should formally address whether, given an arbitrary society of agents, the socially equivalent agent could exist and whether it would be unique.

The distinguishing feature of our approach is the acknowledgement that agent *societies* should be able to self-organise. The realisation of this self-organisation comes in the form of normative prescriptions, and values are the guiding light along the way. This is very much in line with research by Elinor Ostrom and colleagues on communities reliant upon common-pool resources (CPRs) [11]. This type of resources are relatively easy to exploit, and make it very difficult to exclude any individuals from profiting off of its supply. Despite differences in climate, culture and customs, she observed that common-pool resources have a much higher chance of being exploited in a sustainable manner if the community dependent upon them is allowed to craft and enforce their own rules regarding appropriation, maintenance and monitoring activities.

Later on, Ostrom and colleagues identified the common elements and underlying structure of any social interaction, including but not limited to situations related to common-pool resources. They captured their approach in the Institutional Analysis and Development (IAD) framework, which provides analytic tools to study social

interactions at various levels [10, chap. 2]. We would like to situate our proposal in the context of the IAD framework, and particularly the level of analysis where the ethical reasoning happens.

On their surface, social interactions are analysed in operational terms, where agents directly affect their environment through their (possibly joint) actions. One level down in depth are collective-choice situations. It is at this stage where norms affecting the incentive structure of the *operational* level are crafted, proposed, and agreed upon. In our proposal, value-based reasoning happens at this level. Although an agent will reject or accept norms based on its expectation of the effect they will have at the *operational level*, she needs to be situated in a collective-choice arena, i.e. a negotiation process, in order for that reasoning to take place.

The IAD framework establishes two more depth levels below collective-choice situations: the constitutional and meta-constitutional levels. The outcomes from these situations themselves are the rules that establish, for example, the threshold of votes, cast at the collective-choice level, necessary to implement new norms that will shape social interactions at the operational level. Since these levels determine how the negotiation process takes place, they effectively control how the individual values of agents are aggregated. For the outline of the computational ethics model presented here, we will not consider the activities that happen at those levels. It is conceivable, however, that the same norm crafting capabilities that agents employ at the collective choice level could be applied at those deeper levels, in order to change the structure of the norm negotiation process itself.

4 Human and machine values

In summary, the main feature of our proposal is the active role of autonomous agents in the crafting and implementation of technical norms based on their ethical concerns. Does this capability represent a form of moral agency? What is the role of the human designer's morality in all of it? We addressed these questions in the previous sections proposing a model that can be summarised as follows.

First, it should be clearly stated that it is the task of a human team to decide which values are relevant in the multiagent system under design and which form should the goals grounding those values take. Hence, autonomous agents are at all times taking decisions with respect to *human* values. The meaning of values, that is, their manifestation in the particular context of the multiagent system, is an external input subject to human discretion. As far as our proposal goes, autonomous agents do not possess the ability to reason about values as abstract entities, but about the goals they are grounded into and the priorities among those (as set up by the human designer).

It is conceivable to run an evolutionary simulation of agents where agents initially pursue some goal randomly drafted from a pool and reproduce in subsequent generations based on their success. Could this case account for the emergence of values inherent to the agents, or *machine values*, as opposed to human values? We argue not, for the following reason.

There is a big difference between this hypothetical evolutionary approach and the one we have been exposing. Instead of manually selecting the grounding goals

beforehand, the evolutionary approach would set the purpose that values serve in the form of a human-designed fitness function (the bottom right component of Figure 1). Granted, the goals pursued by the most successful individuals would not have been explicitly chosen by an outside human designer. However, the leap from one of the surviving goals into the value that it is grounding is a semantic interpretation that, we conceive, only humans can make. Building the rationale for pursuing a certain goal in terms of higher moral principles is not, as of today, in the reach of autonomous agents. The justifications for the choice of goals that lead to the largest success among agents is still made by humans and in terms of *human* values.

In summary, the values that agents are programmed to promote, even if their grounding goals would be allowed to emerge endogenously, are human values. Even if the moral content that agents work with has its origin outside of themselves, does their capability to select norms, based on their success to promote the instilled values, equate to an ability to embed values into normative systems? According to some philosophers, the capability to embed values is an exclusive competency of agents with moral agency, *i.e.* humans, and inaccessible to software agents [12].

We argue that the technical capacity of artificial agents to select norms based on the values they support does indeed constitute a form of value embedding by the agents into the normative system. According to the Merriam-Webster dictionary, to embed means "to make something an integral part of". It is hard to argue that values are not an integral part of the norms that are eventually enforced, provided that they are precisely chosen based on their alignment with respect to those values. And since agents are capable of "creating norms" (or, at least, searching the space of norms), it is therefore agents who directly embed values into technical normative systems.

Also, this embedding of values into normative systems by the agents is necessarily a collective task. As we have presented in section 3, the negotiations that output the normative systems are a form of social interaction, and the values most supported by the resulting norms do not reflect any individual value preferences but an aggregation over them. So every agent contributes partly, based on the power they hold, to the alignment of the resulting normative system.

So, autonomous agents are capable of embedding values into norms because they have been delegated to do this task. Ultimately, it is a human who generates a representation of values that are an input to the system. In the diagram of Figure 3 (right), the arrows connecting the human to the autonomous agent and the autonomous agent to technical norms should not be seen as independent processes, but as a transmission of values from humans into all components of the multiagent system through the participating agents.

Hence, autonomous agents *can* embed values into a normative system. Is this a form of moral agency? We would argue it is not. Moral agency would require the agents to have a representation of the values themselves, and be able to reason in terms of those abstract concepts. In our proposal, the autonomous agents are provided with the grounding goals as proxies for values, but not with representations of the values themselves. Any moral value under consideration is a very abstract entity, and, as such, is by itself disassociated from any specific instance of behaviour. The association between a specific instance of ethically controversial behaviour and its motivating value, in either direction, is a cognitive ability reserved to humans. Visually, the moral

agency that the human possesses and the autonomous agent lacks resides in the ability to make the connection between the "value" box and the "goal" box in Figure 1.

Some might argue that it would be desirable, and not that technically challenging, to have agents learn the grounding relationships and store it as a set of tree structures, with the values as the root nodes and the context-dependent goals grounding it as the leafs. Intermediate nodes would correspond to various levels of decreasing abstraction, from the root to the leafs. We do not aspire to reach such level of complexity for the time being, and not just because we would like our future work to focus on the role of norms as value-promoting mechanisms. We believe that by letting the designer be the sole responsible for the grounding of values, human control over the system is retained to a much larger degree than if agents were to figure out the grounding goals on their own. This point has potential implications on the legal doctrine on the accountability of autonomous software systems that would require further scrutiny.

5 Summary and future work

In this paper, we have presented and discussed a set of coherent philosophical foundations that, in our view, should underpin a future model of value engineering for autonomous agent systems. The main points shaping our proposal are:

- Values are abstract concepts that, when formalised, are grounded into permanent goals. These goals are programmed into an agent's software by the human designer, and hence agents are instilled with human values.
- Humans have the exclusive competency of grasping values as abstract entities and translating them into the real-world goals that are motivated by that value. Hence, even if agents are endowed with considerable ethically-aware machinery, they still lack the moral agency of humans.
- Technical norms are the primary value-promoting mechanisms. They are intricately related to values by the results that norms are able to achieve and that are compliant with respect to the values of interest.
- It is the group of autonomous agents who directly propose, negotiate over and agree on a new set of norms to be implemented. In that process, agents rely on the value alignment to assess the desirability of any proposal.
- Because all agents actively participate in the generation of technical norms, the negotiation that takes place is a form of aggregation over all the values that the agents regard. The resulting norms do not entirely support any individual agent's values, but the emergent social values of the community.

Obviously, future work should formulate a computational ethics model consistent with this proposal. Despite arguing our position in predominantly philosophical terms, we have provided many hints about the shape that the mathematical formulation should take. We anticipate the norm negotiation component to be the most technically challenging, and expect the choice for value-grounding goals to be the most controversial decision.

References

1. Bostrom, N.: Machine Ethics and Robot Ethics, chap. Ethical Issues in Advanced Artificial Intelligence. Tailor & Francis Group (2017)
2. Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., Ramos, F.: Artificial moral agents: A survey of the current status. Science and Engineering Ethics **26**(2), 501–532 (2019). DOI 10.1007/s11948-019-00151-x
3. Chopra, S., White, L.F.: A Legal Theory for Autonomous Artificial Agents. Michigan University Press (2011)
4. Conger, K., Isaac, M.: Twitter permanently bans trump, capping online revolt. The New York Times (2021). URL https://www.nytimes.com/2021/01/08/technology/twitter-trump-suspended.html
5. Himma, K.E.: Artificial agency, consciousness, and the criteria for moral agency: what properties must an artificial agent have to be a moral agent? Ethics and Information Technology **11**(1), 19–29 (2008). DOI 10.1007/s10676-008-9167-5
6. Kahneman, D.: Thinking, fast and slow. Farrar, Straus and Giroux, New York (2011)
7. Loukides, M.: On computational ethics (2017). URL https://www.oreilly.com/radar/on-computational-ethics/
8. Moor, J.: Is Ethics Computable? Metaphilosophy **26**(1-2), 1–21 (1995). DOI 10.1111/j.1467-9973.1995.tb00553.x
9. Morris-Martin, A., Vos, M.D., Padget, J.: A norm emergence framework for normative mas – position paper (2020)
10. Ostrom, E.: Understanding Institutional Diversity. Princeton University Press (2005)
11. Ostrom, E.: Governing the Commons: The Evolution of Institutions for Collective Action. Canto Classics. Cambridge University Press (2015). DOI 10.1017/CBO9781316423936
12. van de Poel, I.: Embedding values in artificial intelligence (AI) systems. Minds and Machines **30**(3), 385–409 (2020). DOI 10.1007/s11023-020-09537-4
13. Schwartz, S.H.: Universals in the content and structure of values: Theoretical advances and empirical tests in 20 countries. In: Advances in Experimental Social Psychology, pp. 1–65. Elsevier (1992). DOI 10.1016/s0065-2601(08)60281-6
14. Segun, S.T.: From machine ethics to computational ethics. AI & SOCIETY **36**(1), 263–276 (2020). DOI 10.1007/s00146-020-01010-1
15. Serramià, M., López-Sánchez, M., Rodríguez-Aguilar, J.A., Morales, J., Wooldridge, M., Ansotegui, C.: Exploiting moral values to choose the right norms. In: Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society. ACM (2018). DOI 10.1145/3278721.3278735