Metadata of the chapter that will be visualized in SpringerLink

Book Title	Value Engineering in Artificial Intelligence	
Series Title		
Chapter Title	Values, Norms and AI	
Copyright Year	2024	
Copyright HolderName	The Author(s), under exclusive license to Springer Nature Switzerland AG	
Corresponding Author	Family Name	**Steels**
	Particle	
	Given Name	**Luc**
	Prefix	
	Suffix	
	Role	
	Division	
	Organization	Studio Stelluti
	Address	Brussels, Belgium
	Email	steels@arti.vub.ac.be

Abstract	The VALE (value engineering) workshop was held in the econtext of the European Conference on AI (ECAI) in Krakow, Poland on 30 October 2023. This paper briefly summarizes motivations, background concepts, and issues on how to handle moral issues in the use and construction of AI systems.

Values, Norms and AI

Luc Steels(✉)

Studio Stelluti, Brussels, Belgium
Venice International University, Venice, Italy
steels@arti.vub.ac.be

Abstract. The VALE (value engineering) workshop was held in the econtext of the European Conference on AI (ECAI) in Krakow, Poland on 30 October 2023. This paper briefly summarizes motivations, background concepts, and issues on how to handle moral issues in the use and construction of AI systems.

AQ1

1 Motivation

The effectiveness of generative AI to uncannily imitate human intellectual production has triggered a lot of ethical concerns related to: the acquisition (privacy) and use (copyright) of data, the tendency of generative AI to produce inaccurate, invented statements (euphemistically called hallucinations), the abuse of generated output for disinformation, manipulation, cheating and criminal activities, and a negative long term impact on the human capacity for text production and understanding. Also increased automation of processes that have a major impact on human well being and human life using AI (for example automatic weapons or automatic medicine administering) is raising further questions how much control we should leave to machines controlled by AI software.

These concerns have lead to many calls for regulation (including from the companies that produce generative AI), with subsequent initiatives such as the EU AI Act in 2023 and a large increase in funding for the discussion of the ethical issues of AI by social scientists, legal scholars, and policy makers within the EU framework programmes.

But a regulatory approach to AI is in itself not sufficient to achieve trustworthy, safe AI good for humans and society. Technical developments are also needed. If Europe does not invest in this on a much bigger scale than today, it will not have the influence to enact the necessary change. The technical developments should focus on bringing the issue of truth, which informed a lot of work in earlier symbolic AI, back into the picture, but also on how a moral dimension could be more deeply integrated, both in the way AI is used, the way it is designed, and how it operates. The VALE workshop is about the latter: how the moral dimension could be handled by AI systems and their use.

2 The Moral Stance

As philosopher Daniel Dennett pointed out, humans take an intentional stance with respect to other humans (and often towards their pets). The **intentional**

N. Osman and L. Steels (Eds.): VALE 2023, LNCS 14520, pp. 1–7, 2024.
https://doi.org/10.1007/978-3-031-58202-8_1

stance perceives, comprehends and predicts behavior of somebody by assuming that s/he is an agent with beliefs, goals and intentions and interacts with other humans assuming they are also agents that use explicit knowledge of past situations and are capable of deliberation, argumentation, and explanation. One of the central goals of AI is to construct artificial agents that adopt an intentional stance towards their human interlocutors and encourage their users to adopt an intentional stance towards them. This stance is productive for users because the AI agent has such internal complexity that in order to understand and predict (or make strong expectations) about how it will behave, an intentional stance is the most effective way to do so.

A moral stance goes one step further. A **moral stance** perceives, comprehends and predicts the behavior of agents by assuming that they behave according to certain norms that are the reflection of specific values. The moral stance is an extension of the intentional stance. So far the moral stance has not played an important role in the construction of artificial agents or other kinds of applications but it is clear that this is a critical step needed to make AI more acceptable - even though many issues will have to make this possible and even if a moral stance is adopted there are still many issues that remain.

The importance of a moral stance is most obvious in medical domains where norms are explicated in medical protocols. Medical protocols encode practices that are developed, adapted and shared through a social consensus and top-down enforcement (from government, institutions, groups of practitioners). They reflect societal values, not only the rights of individuals but also certain economical considerations or religious beliefs. The design of such medical AI systems therefore has to take into account this moral dimension as well. Moreover an AI system built to support medical decision-making and going beyond the routine application of predefined rules has to do so within the moral bounds expected by their users.

But the moral stance is not only relevant in the medical domain and therefore for the design and usage of applications in that domain. It is present for any kind of application that involves decisions with consequences for human choices, well-being, and/or human rights. For example, we are repelled if a platform, like Youtube, recommends pornography to children, because this conflicts with a key value in our society, namely that children should be protected from sexual exploitation. Although the community guidelines and policies of Youtube, in other words the norms explicitly stated for the use of this platform, forbid this outcome, the platform itself is so far not able to enforce these norms despite significant effort, or it could be that the highest value of owners is to maximise profit rather than protecting children.

3 Norms, Values and Outcomes

There is a consensus in the moral literature that a distinction needs to be made between values and norms. "**Values** are very general, abstract guiding principles that individuals and groups utilise to generate judgements on

a variety of constructs, such as actions, strategies, conventions and policies" [12] Examples of values are: obedience, security, freedom, wealth, forgiveness, care/harm, fairness/cheating, loyalty/betrayal, authority/subversion, and sanctity/degradation, etc. Values are often implicit, resting on common sense 'folk' notions [5]. Some values are sacred, in the sense that those who hold them feel justified in using violence or even giving up their own life, if they consider that their values are not abided by.

"**Norms** (...) establish boundaries, either soft or hard, on individual autonomy through a variety of mechanisms such as social pressure and expectations, constraints on actions, and sanctions for violation or rewards for compliance" [12]. Norms can be implicit, enforced by social pressures and expectations, or they can be explicitly formulated in terms of policies, laws, protocols, usage rules, community rules, etc. The application of norms always implies situation-awareness first. For example in medical end-of-life decisions, it is crucial to get a coherent view of the disease state of the patient and the general context before decisions can be made about treatments.

Norms and values both affect outcomes. **Outcomes** can either be caused by the behavior of an artificial system with respect to a user, for example the behavior of a social robot in a home environment, or the behavior of a user while using an artificial system, for example the behavior of a user on a social media platform who decides to post or propagate certain content. The relations between norms, values and outcomes are summarized in Fig. 1.

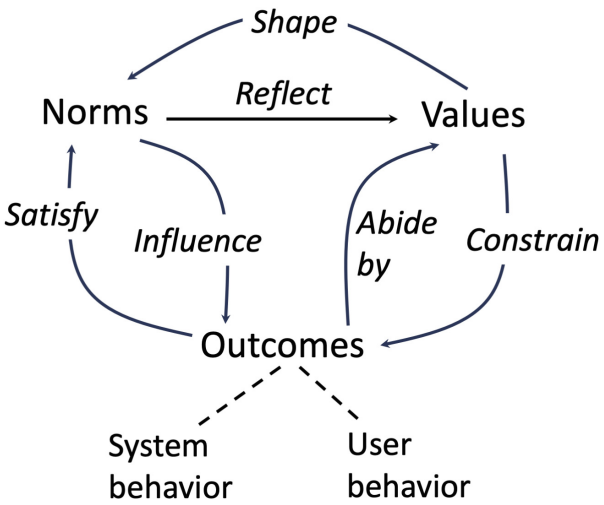

Fig. 1. Norms influence outcomes and outcomes satisfy norms. Values constrain outcomes and outcomes abide by values. Norms reflect values and conversely values shape norms. Outcomes include the effect of the behavior of a system or of a user using a system.

When norms properly reflect values we say that the norms are **value-aligned**. Similarly when the outcomes adequately reflect values we say that these outcomes are **value-aligned**. When outcomes follow norms, we say that the outcomes are **norm-aligned** or **norm-compatible**.

Norms and values are always implicitly part of an artificial system, but they are not always made explicit and if they are, they are usually described in human language and therefore quite vague, ambiguous, and incomplete, relying on the common sense and cultural knowledge of users. Examples are the community rules of social media and their justification in terms of values. Often application designers are not fully aware of what norms and values they implicitly impart to their systems. They follow their intuitive judgements and assume that everyone shares their morality.

Norms may be deliberate or accidental. For example, a trained generative AI model producing text, such as ChatGPT, implicitly reflects the norms and values that are held by the human data being used in training - which may not always be the norms and values that users expect. In this case, a developer can only influence the system by the selection of the data made available. Given the huge amount of data needed, the values and norms adapted by the generative AI model may accidentally be strongly biased or in conflict with societal values [1].

Norms and values can also be made formally and computationally explicit, in other words a system can be given a formal representation of the norms and/or values that determine its behavior or the behavior it expects from its users. In this case we say that the system is **norm and/or value-aware**. Explicit representations of norms are often called policies.

Norm- or value-alignment and norm- or value awareness are not the same thing. You can be aware of a norm, such as that you should not go through a red traffic light, but still do it, perhaps because you see no cars coming and are in a hurry. You are then norm-aware but not norm-aligned. Also your own values may conflict with certain values generally accepted by the people around you and possibly codified as common norms.

One of the big advantages of explicit representations is that outcomes can be explained by the system itself in terms of norms and values. For example, a user may ask why he is blocked on a social media platform and get an answer in terms of which community rules he broke and possibly what values these rules reflect (i.e. why the norm was adopted). A chatbot may refuse to answer a question about a particular topic and can then give an explanation why this is so.

4 AI, Values and Norms

Given these basic concepts, we can now map out the research and development going on to introduce a moral dimension in information systems in general and AI systems in particular. There are activities for each of the bi-directional Norm-Value-Outcome triangle, both for outcomes due to system behavior and for outcomes due to user behavior with a system. In addition there are activities going on to define, formalize and operationalize norms and values. The

VALE workshop can of course not exhaustively cover all these different angles but contains nevertheless a representative subset of current work.

The reports discussed at the workshop have been organized into four major themes:

I. FRAMEWORKS FOR NORMS AND VALUES. This theme addresses what existing or new frameworks for norms and values have been proposed and how they can be represented in AI systems in order to make them value-aware. There are reports on a new 'contractual' framework linking moral decisions with the justifications of actions towards others [11], a way to capture the common sense assumptions that are often left implicit in formalisations of norms and values, [5] a method how to integrate normative reasoning into argumentation [16] and a framework that takes the perspective of others into account [12]. This section also reports on how to acquire norms and values through natural language analysis [5].

II. DETECTION OF MORAL VALUES: This theme addresses how to figure out what the norms and values are of a system or of users of a system and whether these are aligned with human values. There is first a report on how the moral values in textual posts on social media can be categorized in terms of a moral stance, [3] then two reports on how the implicit norms and values in responses by generative AI can be queried to see whether they are aligned with human norms and values, [1] and [4], and a report on detecting moral values in political argumentations. [14].

III. LEARNING AND ENGINEERING OF POLICIES: This theme is about how policies for implementing norms and values can be acquired or designed. There are reports on the learning of value-aligned policies [8], methods to use a principled logic-based design approach [13], and numerical methods to derive equations and parameters for computing value alignment [9,10].

IV. IMPLEMENTATION OF NORMS: The final theme is about how norms can be implemented in concrete AI applications. There are reports on applications in military decision-making [15], taxi scheduling [7], employee hiring [6], and school placement [2]. Each of these applications also discusses fundamental issues about the representation of values and norms, how alignment is established, and how values and norms can be acquired. The papers in other themes also address various applications, specifically for classification of social media posts [3], social robots [1], tax payment: [9], and management of common pool resources [10] (agriculture), [8] (water distribution).

Besides tackling many issues within a rich variety of application domains, we see across the different papers significant variation in the AI methods that are being deployed, including Answer-set programming [2,6,10], Multi-layered inductive neural learning: [1,4], Constrained reinforcement learning: [8], Knowledge graphs: [5], and Numerical modeling: [9,10,15].

5 Issues

The introduction of a moral stance in the design, implementation and use of AI systems is certainly a step forward towards a safer and human-friendly AI, although many hurdles remain to be overcome before the proposed methods can be applied in a routine manner. But there are also dangers that we need to worry about. Perhaps the biggest one is an increase in technocratic control, which is already a big problem in contemporary society.

Values often arise because they make sense for a specific community at a particular point in time but they may linger on, even if the societal or ecological conditions have changed and they are no longer objectively justifiable. Also norms based on values may require adaptation and flexibility. The laws made by parliaments therefore deliberately remain partially ambiguous and underspecified so that they can be flexibly applied by courts, allowed to evolve as societal conditions change, and make it possible to deal with outliers.

At the moment the values and norms underlying the outcomes of AI systems (or the human moderators behind those systems) are not explicit and in some cases change at the whim of system owners or change imperceptibly with the usage of more data for training. This contrasts to legally enshrined norms that go through a careful process of vetting and societal approval. Introducing explicit representations of values and norms and making systems accountable in the sense of able to explain the moral foundation of their own actions is a step in the right direction. Also explanations how and why inappropriate action of their users are constrained is certainly positive.

At the same time we need to worry much more how we can retain the possibility of flexibility and adaptivity and how we can respect the privacy and self-determination of the individual properly. Until that is done we should refrain from too hastily introducing value-aware AI in real world settings.

Acknowledgement. The writing of this paper with funding through the EU pathfinder VALAWAI (Value-aware AI) project to Studio Stelluti and the EU H2020 MUHAI project to the Venice International University. I am indebted to other partners in the VALAWAI project, particularly Nardine Osman and Carles Sierra from IIIA in Barcelona, Oscar Vilarroya, Clara Pretus and Luis Marcos from IMIM in Barcelona, and Giulio Prevedello from the Sony Computer Science Laboratory in Paris, for discussions related to values, norms and AI.

References

1. Abbo, G.A., Marchesi, S., Belpaeme, T., Wykowska, A.: Do LLMs show traits of value awareness? (2023, this volume)
2. Arias, J., Rebato, M.M., Rodriguez, J.A., Ossowski, S.: Value awareness and process automation: a reflection through school place allocation models (2023, this volume)
3. Brugnoli, E., Gravino, P., Prevedello, G.: Moral values in social media for disinformation and hate (2023, this volume)

4. Bulla, L., Mongiovi, M., Gangemi, A.: Do language models understand morality? Towards a robust detection of the moral content? (2023, this volume)
5. De Giorgis, S., Gangemi, A.: That's all folks: a KG of values as commonsense social (2023, this volume)
6. Fernández-Martinez, C., Fernandez, A., Arias, J.: Value-based reasoning scenario in employee hiring and onboarding using answer set programming (2023, this volume)
7. Garcia, M., Cordova, C., Taverner, J., Palanca, J., del Val, E., Argente, E.: Towards a distributed platform for normative reasoning and value alignment in multi-agent systems (2023, this volume)
8. Holgado, A., Arias, J., Billhardt, H., Ossowskia, S.: Algorithms for learning value-aligned policies considering admissibility relaxation (2023, this volume)
9. Karanik, M., Billhardt, H., Fernández, A., Ossowski, S.: Exploiting value system structure for value-aligned decision-making (2023, this volume)
10. Lujak, M., Fernandez, A., Billhardt, H., Ossowski, S., Herrero, J.A., Sánchez, A.L.: Exploiting value system structure for value-aligned decision-making (2023, this volume)
11. Marcos, L., Marchesi, S., Wykowska, A., Pretus, C.: Moral agents as relational systems: the contract-based model of moral cognition for AI (2023, this volume)
12. Montes, N., Osman, N., Sierra, C.: Perspective-dependent value alignment of norms (2023, this volume)
13. Noriega, P., Plaza, E.: An AGV approach to value engineering, and beyond (2023, this volume)
14. Zhang, H., Landowska, A., Budzynska, K.: Detection and analysis of moral values in argumentation (2023, this volume)
15. Zurek, T., van Engers, T., Kwik, J.: Values, proportionality, and uncertainty in military autonomous devices (2023, this volume)
16. Zurek, T., Wyner, A.: Towards a formalisation of motivated reasoning and the roots of conflict (2023, this volume)

Author Queries

Chapter 1

Query Refs.	Details Required	Author's response
AQ1	This is to inform you that corresponding author email address has been identified as per the information available in the Copyright form.	