Ethics by Design

Necessity or Curse?

Dignum, Virginia; Baldoni, Matteo; Baroglio, Cristina; Caon, Maurizio; Chatila, Raja; Dennis, Louise; Génova, Gonzalo; Kließ, Malte S.; De Wildt, Tristan; More Authors

DOI
10.1145/3278721.3278745

Publication date
2018

Published in
AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society

Ethics by Design: Necessity or Curse?

Virginia Dignum
UmeåUniversity
Umeå, Sweden

Matteo Baldoni
Università degli Studi di Torino
Torino, Italy

Cristina Baroglio
Università degli Studi di Torino
Torino, Italy

Maurizio Caon
University of Applied Sciences and Arts Western Switzerland
Fribourg, Switzerland

Raja Chatila
Institut des Systèmes Intelligents et de Robotique, Sorbonne Université
Paris, France

Louise Dennis*
University of Liverpool
Liverpool, UK

Gonzalo Génova
Universidad Carlos III de Madrid
Leganés, Spain

Galit Haim
The College of Management Academic Studies
Rishon Lezion, Israel

Malte S. Kließ
Delft Technical University
Delft, The Netherlands

Maite Lopez-Sanchez
Universitat de Barcelona
Barcelona, Spain

Roberto Micalizio
Università degli Studi di Torino
Torino, Italy

Juan Pavón
Universidad Complutense Madrid
Madrid, Spain

Marija Slavkovik
University of Bergen
Bergen, Norway

Matthijs Smakman
HU University of Applied Sciences Utrecht
Utrecht, The Netherlands

Marlies van Steenbergen
HU University of Applied Sciences Utrecht
Utrecht, The Netherlands

Stefano Tedeschi
Università degli Studi di Torino
Torino, Italy

Leon van der Torre
University of Luxembourg
Luxembourg, Luxembourg

Serena Villata
Université Côte d'Azur, CNRS, Inria
Sophia Antipolis, France

Tristan de Wildt
Delft Technical University
Delft, The Netherlands

ABSTRACT

Ethics by Design concerns the methods, algorithms and tools needed to endow autonomous agents with the capability to reason about the ethical aspects of their decisions, and the methods, tools and formalisms to guarantee that an agent's behavior remains within given moral bounds. In this context some questions arise: How and to what extent can agents understand the social reality in which they operate, and the other intelligences (AI, animals and humans) with which they co-exist? What are the ethical concerns in the emerging new forms of society, and how do we ensure the human dimension is upheld in interactions and decisions by autonomous agents?. But overall, the central question is: *"Can we, and should we, build ethically-aware agents?"*

This paper presents initial conclusions from the thematic day of the same name held at PRIMA2017, on October 2017.[1]

CCS CONCEPTS

• **Computing methodologies** → **Multi-agent systems**; *Philosophical/theoretical foundations of artificial intelligence*;

KEYWORDS

ACM proceedings, LaTeX, text tagging

ACM Reference Format:
Virginia Dignum, Matteo Baldoni, Cristina Baroglio, Maurizio Caon, Raja Chatila, Louise Dennis, Gonzalo Génova, Galit Haim, Malte S. Kließ, Maite Lopez-Sanchez, Roberto Micalizio, Juan Pavón, Marija Slavkovik, Matthijs Smakman, Marlies van Steenbergen, Stefano Tedeschi, Leon van der Torre, Serena Villata, and Tristan de Wildt. 2018. Ethics by Design: Necessity or Curse?. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3278721.3278745

*Corresponding author.

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02.
https://doi.org/10.1145/3278721.3278745

[1] https://prima2017.gforge.uni.lu/ethics.html

1 INTRODUCTION

In the near future, we will all experience Artificial Intelligence (AI) applications making decisions and acting in our world, with a greater level of autonomy, in many areas of application, including domains such as transportation, finance, health-care, education, public safety and security, and entertainment. However, to fully benefit from the potential of AI, we need more than improved perception and search algorithms and increased computational power or solving capabilities. We need to make sure that these technologies are aligned with our moral values and ethical principles. That is, AI will have to behave in a way that is beneficial to people beyond reaching functional goals and addressing technical problems. This is necessary to ensure an elevated level of trust between humans and technology, which is needed for a fruitful pervasive use of AI in our daily lives.

Ethical, legal and societal (ELS) issues raised by the development of Artificial Intelligence, Robotics and Autonomous Systems have recently generated strong interest both among the general public and in the involved scientific communities, with the development of applications often based on deep learning programs that are prone to bias, the wide exploitation of personal data, or new applications and use cases, such as personal robotics, autonomous cars or autonomous weapons. These ELS questions cover a wide range of issues such as: the future of employment, privacy and data protection, surveillance, interaction with vulnerable people, human dignity, autonomous decision-making, the moral responsibility and legal liability of robots, imitation of living beings and humans, human augmentation, and the status of robots in society.

In fact, the issue of the ethical aspects of AI is hot: you cannot click on a news site nor open a newspaper without finding an article about the role of ethics in AI.[2] If we wish to avoid unintended negative consequences for society, the hype around this subject is warranted [15]. However, we need to go beyond the hype and start taking decisions about responsibility for AI behavior and its impact on society. Several initiatives are already analyzing this issue, including amongst others the IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems,the Partnership on AI, and the AI Now Institute. At the level of national governments, many countries are also looking at means to regulate AI. In all cases, the alternatives being considered can be divided into two types: *regulation* by means of legislation and standards, or *design*, ensuring that the systems themselves take ethical decisions at all times. This paper is concerned with the latter.

Ethics by Design concerns the methods, algorithms and tools needed to endow autonomous agents with the capability to reason about the ethical aspects of their decisions, and the methods, tools and formalisms to guarantee that an agent's behavior remains within given moral bounds. How and to what extent can agents understand the social reality in which they operate, and the other intelligences (AI, animal and humans) with which it co-exists? What are the ethical concerns in the emerging new forms of society, and how do we ensure the human dimension is upheld in interactions and decisions by autonomous agents.

The central question is, therefore:

Can we, and should we, build ethically-aware agents?

This paper presents initial conclusions from the thematic day of the same name held at PRIMA2017, on October 2017[1]. It is organized as follows. In the next section, we discuss the issue of responsibility in AI. The following section, presents initial considerations on the means and rationale for guaranteeing ethical behavior 'by design', i.e. embedded in the system's implementation. Next, we present a few illustrative examples of ethics by design. Finally, we present our conclusions and directions for future work.

2 ETHICS AND RESPONSIBILITY IN AI

As advances in AI occur at high speed, many questions arise about social, economic, political, technological, legal, ethical and philosophical issues. Can machines make moral decisions? Should artificial systems ever be treated as ethical entities? What are the legal and ethical consequences of human enhancement technologies, or cyber-genetic technologies? What are the consequences of extended government, corporate, and other organisational access to knowledge and predictions concerning citizen behaviour? How can moral, societal and legal values be part of the design process? How and when should governments and the general public intervene?

Answering these and related questions requires a whole new understanding of Ethics with respect to control and autonomy, in the changing socio-technical reality. The urgency of these issues is acknowledged by researchers and policy makers alike. Moreover, implementing ethical actions in machines will help us better understand ethics overall. To enable the required technological developments and responses, AI researchers and practitioners will need to be able to take moral, societal and legal values into account in the design of AI systems. AI requires researchers who can elicit and represent human values, translate these values into technical requirements, innovate in cases of moral overload when numerous values are to be incorporated, and who can demonstrate that design solutions realize the values wished for.

At the same time, considering the ethical and societal consequences of actions and decisions by AI systems requires a mental shift from researchers and developers towards the goal of ensuring trust rather than focusing on performance alone. This shift will lead to novel and exciting techniques and applications, and will open up a new direction in AI research. Current development of AI algorithms has so far been led by the goal of improving performance, leading to efficient but very opaque algorithms. Developing methods to inspect algorithms and their results, and to question the system about its reasoning should be a priority in AI.

It has been argued [22] that in a range of domains, a key factor in humans willing to trust autonomous systems is that the systems need to be able to *explain* why they performed a certain course of action. For example, the IEEE Ethically Aligned Design report suggests "*... a why-did-you-do-that button which, when pressed, causes the robot to explain the action it just took*" [18, page 20]. The importance of explanation has also been recognized by the European Union, and is captured in the EU General Data Protection Regulation[3]

Moreover, putting human values at the core of AI systems calls for transparent data governance mechanisms to ensure that data used to train algorithms and guide decision-making is collected,

[2] E.g., https://goo.gl/ZEdU9H and https://goo.gl/JyrY5s.

[3] http://tinyurl.com/GDPREU2016

created, and managed in a fair and clear manner, taking care to minimize bias and enforce privacy and security. New and more ambitious forms of governance is one of the most pressing needs for ensuring that inevitable AI advances will serve societal good.

If we want to ensure that AI-related developments are to be for societal good there are three aspects of particular concern [4]. These are the principles of Accountability, Responsibility and Transparency (ART):

- **Accountability** refers to a system's need to explain and justify decisions and actions to its partners, users and others with whom it interacts. To ensure accountability, decisions must be derivable from, and explained by, the decision-making algorithms used. This includes the need for representation of the moral values and societal norms holding in the context of operation, which the agent uses for deliberation. Accountability in AI requires both functionality for guiding action (by forming beliefs and making decisions), and for explanation (by placing decisions in a broader context and by classifying them along moral values).
- **Responsibility** refers both to the capability of AI systems and to the role of people interacting with it. Both need to be considered when accounting for a decision and when diagnosing errors or unexpected results. As the chain of responsibility grows, means are needed to link the AI system's decisions to the fair use of data and to the actions of stakeholders involved in the system's decision.
- **Transparency** refers to the need to describe, inspect and reproduce the mechanisms through which an AI system makes decisions and learns to adapt to its environment, and to the governance of the data used and created. Current AI algorithms are essentially black boxes. However, regulators and users demand explanation and clarity. Methods are needed to inspect algorithms and their results and to manage data provenance and dynamics.

How AI systems comply to these principles, depends on how ethical responsibility is being considered. In the case of regulation, ART is placed with the people and the institutions that define and monitor the behavior of the system. If we assume an AI system has no capacity for ethical reasoning then we have a limited number of approaches. The system could have a human supervisor. This requires the inclusion of means to ensure shared awareness of a situation, so the supervisor has enough information to determine if intervention is necessary. Such interactive control systems are known as human-in-the-loop control systems [8]. Alternatively, the environment can be regulated in such a way that deviation is impossible, and therefore moral decisions by the autonomous system are not needed. This is the mechanism used in smart highways, linking road vehicles to their physical surroundings, where the road infrastructure controls the vehicles [12].

In the case of Ethics by Design, the AI system is the ethical agent itself. These systems are also known as *Artificial Moral Agents (AMA)*, i.e. AI systems able to incorporate moral reasoning in their deliberation and to explain their behavior in terms of moral concepts. An AMA [21] can autonomously evaluate the moral and societal consequences of its decisions and use this evaluation in their decision-making process. Here moral refers to principles regarding *right* and *wrong*, and explanation refers to algorithmic mechanisms to provide a qualitative understanding of the relationship between the system's beliefs and its decisions. This approach requires complex decision making algorithms, based e.g. on deontic logics, which may require a mixture of top-down explicit design and bottom up derivation e.g. based on reinforcement learning.

3 ISSUES ON ETHICS BY DESIGN

Ensuring Ethics by Design raises many questions, not only related to its feasibility but also to its desirability. In fact, the decision to have artifacts taking ethical action, is in itself an ethical decision. In this section, we try to answer a few of these questions, namely understand the meaning of responsibility when the responsible actor is an artifact, the issue of determining the right norms to implement, and how to implement them, and finally, the meaning of artificial ethical decisions.

3.1 Ethical decision by AI systems

Decision-making as a process has been studied extensively in many disciplines of social sciences, mathematics and psychology. Both the descriptive aspects, how people make decisions, and the prescriptive aspects, namely methods for making decisions are a continuous object of scientific interests. Most research addresses how humans make or should make decisions. Within AI we need to consider how to build an algorithm that discerns between an ethical option and an unethical one. We also need to consider the ethical impact of autonomous decision-making algorithms in general. Given that these aspects of decision making have been far more extensively studied for human decision-makers, the question is, is there a difference between a (ethical) decision made by a person and the same decision made by a machine? The answer to this question is positive, and we will try to elaborate some of the points of difference.

One of the differences between ethical decisions made by humans and those made by machines is in the evaluation of the decision-making process. People are liable to themselves and to society for the morality of their choices and actions [6]. An evaluation of the decision-making process is done, if at all, *ex post*. People are rarely asked to explain their decision making process, and when this is done, it is done with the purpose of assigning liability. However, the explainability and justification of a machine's decision should be routine and can leads to improvement of the system as a whole particularly when a decision is shown to be wrong. The need for explainability makes the use of any kind of learning in order to derive ethical behaviour particularly challenging – for the forseeable abstract concepts such as dignity can not be taught nor understood by machines. Explainable AI should take account of the work that has been done on human explanation of behavior. See [11] for a good survey.

Another difference between machines and people is that the latter are assumed, by default, to be moral agents unless information exists to demonstrate the contrary. With machines no such assumption can be made. If anything, machines are assumed to be incapable of moral reasoning. Society must therefore require a proof, or certification, for the ethical reasoning abilities of a machine, but we do not have any clear description, let alone consensus

on the nature of that proof. People are often able to extrapolate what is the right thing to do by considering a very small number of examples. However, complex ethical decisions cannot be made case by case by a machine, given the current capabilities of Artificial Intelligence. We should not assume this situation will change in the near future – ethical decision making by machines will, for some time, be guided heavily by general rules and the nature of these rules should form the basis of any certification process.

While to err is understood to be human, it is unclear how tolerant society is to machines making the wrong decisions. It appears that machine decision-making is held to a higher standard than human decision-making [10]. One reason for this could be that certain justifications for making a decision, such as empathy, feeling distracted or confused, are only valid arguments or "excuses" for people and do not apply to machines. In any case, the attribution of some accountability to machines, implying the necessity for transparency and explanation, should never replace human responsibility nor be used as a means to release people from liability. Means to link the machine's actions and decisions to manufacturer, designer, owner, user are needed in order to enable the sharing of liability for the machine's decisions.

These considerations mean that before we can begin designing ethical reasoning into machines we need to develop a clearer understanding of the nature of machine ethics.

Lastly we observe that while all decisions can be seen as ethical when viewed from an appropriate level of abstraction (in that any action taken by any computational system will at the very least use up some resource, and will reflect (and potentially help shape) the priorities and values of the system's designer, programmer or user community), ethics is not at the forefront of most decision making at a practical level. We suggest that ethical decisions are those decisions that are related to, or directly impact upon, human dignity and well-being. In general we would anticipate that explicitly ethics based reasoning in machines will only be required in specialised circumstances, either because the system has had to reason outside some limits pre-determined as ethical (e.g., [3]) or because a conflict between ethical principles has been detected and requires resolution.

This means that while we may draw upon existing techniques for designing and implementing reasoning in machines in order to implement ethical machine reasoning, we need to be aware of the different nature and context of ethical reasoning, and explicitly include this awareness in our design and implementation.

In summary, the following issues illustrate the main differences between the ethics of human decision and those of machines:

- Ethical machine reasoning requires explainability, probably in terms of abstract concepts such as human well-being and dignity.
- Machines can only select within a bounded set of categories or decisions provided to them directly or indirectly (e.g., learning) by a human programmer. This contrasts to the case-by-case reasoning used by humans tasked with making complex ethical decisions.
- Accountability must remain on the humans – those who designed or programmed the machine, or those who customised and deployed it, or those who use it.
- Ethical machine reasoning has distinct features that distinguish it from ethical human reasoning and goal-based machine reasoning.

3.2 AI responsibility

Responsibility is core to AI development. Responsibility refers to the role of people as they develop, manufacture, sell, and use AI systems, but also to the capability of AI systems to answer for their decisions and identify errors or unexpected results. As the chain of responsibility grows, means are needed to link the AI system's decisions to the fair use of data and to the actions of stakeholders involved in the system's decision. Which means that AI reasoning should link moral, societal, and legal values, and that it should be capable of being questioned.

AI systems are increasing in autonomy, which includes the capability to take decisions by themselves, but the discussion of who is responsible for these decisions is not yet sufficiently determined. In general machine reasoning has worked best where there are a limited number (preferably one) of clearly stated objectives and any ambiguity is around the best way to achieve the objective, not upon deciding which objectives are more important than which others and to what degree. Moreover, particularly when conflicts arise, it may be necessary to make a decision even when no solution (in the form of a decision that violates no ethical principle) exists.

The concept of responsibility is related to liability. Liability is a typically human concept, related to others like blame and sanctioning. This has important drawbacks when shifting to the field of software systems. Simply put, a computer cannot be fined or put in jail when a bad decision is made. However, these decisions can have severe social implications. For this reason other protection mechanisms must be developed and enforced, guaranteeing a sort of transfer of responsibility from the machine to some other socially significant entity.

The responsibility for implementing the right behavior is not only a role for software developers but for all stakeholders. Even though the implementation must comply with well-defined requirements and constraints, results are not the sole responsibility of the developers, and certainly not of the machine, but are shared by different stakeholders. The main problem, here, is that the decision-making process carried out by an intelligent system, especially a learning one, cannot be inspected, in general, as it could be done with a traditional algorithm. At the same time, often these kinds of systems offer no explanation about a given decision. The system becomes, then, a sort of black-box ifrom the point of view of the developers as well as the users.

In this setting, institutions such as governments play a central role since they embody and reify the socio-cultural context in which ethical decisions acquire their meaning. They have, then, the big responsibility of clearly setting up the ethical framework specification to which the given intelligent system should comply. Significant in this perspective is the recent creation of several frameworks for accountability and responsibility in the main intergovernmental organizations, such as the United Nations. Technological companies which develop and distribute intelligent systems must ensure that these specifications are respected and put in place appropriate countermeasures against the event that the machine does something

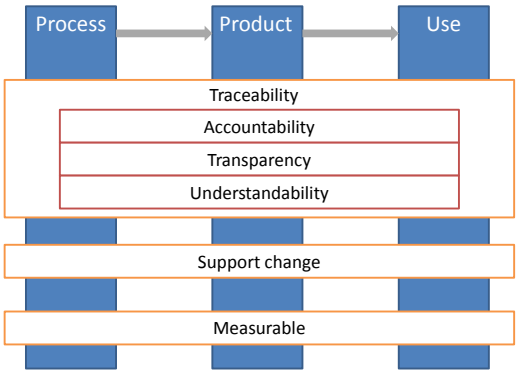

Figure 1: Ethical aspects of process, product and use.

wrong, i.e. it does not fully comply to the specification. Since the decisions made by are becoming more and more critical, the recovery mechanism must be effective and responsive, as well. Developers, in turn, have the responsibility to find practical mechanisms to ensure that these black-box systems respect important properties, despite their learning and decision-making processes. Last, but not least, end users bring also have responsibilities. The misuse of an intelligent system, such as a self-driving car, could lead to unpredictable behaviour and to fatal outcomes.

In the next section we address how the chain of responsibility is established.

3.3 Norms and their implementation

The specification of norms of behaviour is central to the development of artificial agents that can behave in ethically correct ways. I.e. the agent must have some representation of what constitutes 'normal', or acceptable, behavior, and have the means to act accordingly. However, even though the development of normative agents is a very active research fields in AI (cf. [2, 7], the concept of norm is commonly introduced in the AI literature without much discussion. Most importantly, most of this work assumes that the set of norms to be given is fixed. However, current technological developments are changing societal norms. Therefore, it is necessary to develop methods and tools that not only enable the representation of norms into AI architectures, but most importantly, support norm elicitation, norm change and can dynamically adapt to new situations. Work in Design for Values methods can be useful in that it makes norms and values explicit, and provides a means to link values to norms to implemented functionalities in context sensitive ways [5, 19, 20].

Figure 1 depicts some of the issues that should be considered when working on norms at three different stages: Process, Product, and Usage. Process refers to the method used for norm creation. Product stands for the characteristics of resulting norms. Usage means actual norm implementation.

Along these subsequent stages there are three issues that we think are key for the success of the overall regulatory process:

(1) Traceability must be guaranteed along all the stages. Traceability can be ensured in terms of accountability, transparency, and understandability;
(2) Support to norm change is vital for guaranteeing the success of the overall regulatory process, since most regulated systems are dynamic, and, therefore, regulations must adapt accordingly, and,
(3) Measurability will provide the objective means to assess the quality of the different stages.

It is worth noticing that handling these issues may imply specific interventions at different stages, as it is the case in software development, where the development process is subject to specifications which differ from the quality insurance measures related to the resulting product.

4 ILLUSTRATIONS OF ETHICS BY DESIGN

During the workshop, the issue of Ethics by Design was illustrated from different perspectives, ranging from its philosophical grounding, to methodological issues, to concrete practical applications. In this section, we briefly present these different contributions.

Raja Chatila explained that Ethics by Design are sometimes related to classical issues in ethical philosophy and law by transposing them to intelligent machines, but they also pose new problems on which reflection must mobilize interdisciplinary communities in order to grasp globally the scientific, technical, and social aspects. The question in developing theses technologies, which might have an unprecedented impact on our society, is finally about how to make them aligned with the values on which human rights and well-being are based.

From the perspective of the designers of such systems, two main issues are central:

(1) research methodologies and design processes themselves: how to define and adopt an ethical and responsible methodology for developing these technological systems so that they are transparent, explainable and so that they comply with human values? This involves several aspects that transform product lifecycle management approaches;
(2) when decisions are delegated to so-called autonomous systems, is it possible to embed ethical reasoning in their decision-making processes?

Maite Lopez-Sanchez discussed on values and norms. Moral values and norms are deeply rooted in most societies. Values –such as equality or respect– are often considered as moral standards for distinguishing between right and wrong (i.e., good or evil). As for norms, they constitute coordination mechanisms that govern the (expected) behaviour in specific situations —such as waiting for turn in queues. Although both guide our conduct, values are more general than norms. In fact, inspired by [1], we consider that norms promote moral values. Thus, for example, a rule 'wear dark at funerals' regulates a specific situation that promotes respect. When deciding upon the norms to enact in a society, a question naturally rises: How to choose the "right" norms? We argue that moral value promotion can be used as a criteria in this norm decision making process. In fact, from the specification of both norm value promotion and shared preferences over moral values, we can encode, as in [9], an optimisation problem as a linear program that can be automatically solved with state-of-the-art solvers.

Juan Pavón presented, from the point of view of Responsible Research and Innovation (RRI), the need to consider the involvement

of different stakeholders along the whole lifecycle of the intelligent systems. This implies the support for new methodologies that cope with the issues that require their cooperation. Some of these issues are the use of different languages and concepts, local vs. global perspectives, social vs. individual preferences, assignment of responsibilities (legal, specification, design, implementation, testing, acceptance, usage), etc.

Gonzalo Génova addressed the question of whether we can teach ethics to machines. Undoubtedly, we can program machines to behave according to ethical principles. We have been doing this for decades: every time a machine chooses among alternate courses of action, it does so based on (implicit or explicit) ethical principles that guide the evaluation of alternatives. In this sense, we can say that the designer has "taught" ethics to the machine. A more interesting situation is posed when we use AI techniques to make the machine learn due behavior by means of labeled examples. The ethical principles/rules need not be explicit, and the machine is able to extract and learn them "by itself". This is the approach followed by the MIT Moral Machine. However, these approaches have a strong bias to the consideration of ethical behavior as fundamentally consisting in following a code of conduct: either a code that is explicitly programmed, or one that is inferred from ethically labeled samples. The more fundamental issue of recognizing ethical values (and valuable entities) is left out, an issue that leads us back to the Turing Test: how can a machine recognize a human being, understood as a being worthy of respect?

Marija Slavkovik discussed on Algorithms and Autonomy. It is easy to see how an autonomous agent that shares the physical world with us can have an ethical impact on our society, or be an agent of unethical behaviour. A web based service can offer far more options than a rational user is able and willing to consider in order to decide what to read, consume, or purchase. Some preprocessing of choices is therefore necessary. Big data analysis combined with user tracking have made it possible to curtail options in a tailor made fashion. Intelligent agents dynamically determine what options are most relevant for people like you when you access a service. However you, as a user, do not get to approve or even see the parameters that are used to define you and that directly determine which choices are shown to you. Issues of privacy violation during user tracking have been extensively discussed and some legislation has been put in motion to deter the grossest of violations. Many questions remain untackled: Does this constitute a violation of user autonomy? By selecting the options available for a particular group of people, can a strategic designer design group behaviour?

Marlies van Steenbergen put the focus on the ethics of the design process itself. This concerns both the responsible use of personal data and the transparency of AI services. To this effect, the fields of *value sensitive design* [5, 20] and *choice architecture* [17] provide concrete design guidelines on how to build 'ethics' into the design and development of digital services Questions regarding ethics in digitalization must be addressed in a multidimensional and multidisciplinary manner.

Maurizio Caon considered the issues in health monitoring systems. In the near future, systems and sensors will be so pervasive that they will be able to monitor the users ubiquitously and constantly. These systems will be able to accurately assess the users' behaviors and physiological parameters. How ethical is it to influence the user's behavior? These systems will be always active. What happens when a user deliberately decides to perform a "deprecated" activity? Will it become a moral fault not willing to conform to standard behaviors? Will it be a moral fault not wanting to conform to medical standards for prevention although this can compromise pleasure and eventually a general well-being? Several national health systems (NHS) are already working on the development of digital personal health folders [14]. When these systems are pervasive, will the NHS be allowed to store these data? Will it be allowed to tax bad behaviors? Can the NHSs share these data with insurance companies? Can these companies vary their insurance policy prices based on the monitored behaviors? Finally, how is it possible to design systems that can take into account all these ethical issues?

Matthijs Smakman discussed on Robot Tutors and Moral Concepts. Robots are entering into our daily lives and becoming more social. One example is the tutor robot, a promising new technology that is expected to support teachers and improve learning. However, a moral framework for applying tutor robots in a justified way is still missing.

Moral concepts such as morally right, moral obligation and fairness, are deeply rooted in society, and, among other things, used to motivate and reinforce behaviour towards what promotes well-being. However, it is questionable whether robots need these concepts to guide their behaviour, especially since they can be programmed as completely rational agents. This would entail that robots don't need these additional motivations, and they could suffice with a concept of what promotes well-being.

Tristan de Wildt considered the case of energy systems. The urgent need for the energy transition and the fact that infrastructural changes in the energy sector often lead to acceptance issues from the public, new energy systems need to be designed in a more acceptable way. The evaluation of the fulfillment of values by specific business models can be done using agent-based modelling and the capability approach [13, 16]. In these models, heterogeneous agents interact with each other to evaluate the capabilities that they have to achieve certain actions allowing them to fulfill certain values. A design is hence considered more acceptable if opportunities to achieve relevant values are safeguarded.

5 CONCLUSIONS

With the aim of opening paths to future research, this paper highlights some of the issues that arise when considering ethical aspects in the design of autonomous systems. Despite the high variety of issues, first future steps may include: the identification and connection of interested partners, ideally from academy, industry, and social organisations; the clarification and dissemination of machine ethics specificities, that distinguish the field from IT ethics or general Artificial Intelligence; and the elaboration of a position outline for machine ethics programmers.

ACKNOWLEDGMENTS

Louise Dennis was supported by the EPSRC "Verifiable Autonomy" research project (EP/L024845), Juan Pavón by H2020 FoTRRIS project (665906), and Maite Lopez-Sanchez by the Spanish research project "Collectiveware" (TIN2015-66863-C2-1-R).

REFERENCES

[1] Trevor J. M. Bench-Capon and Katie Atkinson. 2009. Abstract Argumentation and Values. In *Argumentation in Artificial Intelligence*. 45–64. https://doi.org/10.1007/978-0-387-98197-0_3

[2] Jan Broersen, Mehdi Dastani, Joris Hulstijn, Zisheng Huang, and Leendert van der Torre. 2001. The BOID Architecture: Conflicts Between Beliefs, Obligations, Intentions and Desires. In *Proceedings of the Fifth International Conference on Autonomous Agents (AGENTS '01)*. ACM, New York, NY, USA, 9–16. https://doi.org/10.1145/375735.375766

[3] L. Dennis, M. Fisher, M. Slavkovik, and M. Webster. 2016. Formal verification of ethical choices in autonomous systems. *Robotics and Autonomous Systems* 77 (2016), 1–14. https://doi.org/10.1016/j.robot.2015.11.012

[4] Virginia Dignum. 2017. Responsible Autonomy. In *Proceedings of IJCAI 2017*. 4698–4704.

[5] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[6] B. Gert and J. Gert. 2017. The Definition of Morality. In *The Stanford Encyclopedia of Philosophy* (fall 2017 ed.), E.N. Zalta (Ed.). Metaphysics Research Lab, Stanford University.

[7] Christopher D Hollander and Annie S Wu. 2011. The current state of normative agent-based systems. *Journal of Artificial Societies and Social Simulation* 14, 2 (2011), 6.

[8] Wenchao Li, Dorsa Sadigh, Shankar Sastry, and Sanjit Seshia. 2014. *Synthesis for Human-in-the-Loop Control Systems*. Springer, 470–484. https://doi.org/10.1007/978-3-642-54862-8_40

[9] Maite Lopez-Sanchez, Marc Serramia, Juan A. Rodríguez-Aguilar, Javier Morales, and Michael Wooldridge. 2017. Automating decision making to help establish norm-based regulations. In *Proceedings of the 16th Conference on Autonomous Agents and MultiAgent Systems*. 1613–1615.

[10] B. F. Malle, M. Scheutz, T. Arnold, J. Voiklis, and C. Cusimano. 2015. Sacrifice One For the Good of Many?: People Apply Different Moral Norms to Human and Robot Agents. In *Proceedings of the Tenth Annual ACM/IEEE International Conference on Human-Robot Interaction (HRI '15)*. ACM, 117–124. https://doi.org/10.1145/2696454.2696458

[11] Tim Miller. 2017. Explanation in Artificial Intelligence: Insights from the Social Sciences. *CoRR* abs/1706.07269 (2017). http://arxiv.org/abs/1706.07269

[12] James Misener and S. Shladover. 2006. PATH investigations in vehicle-roadside cooperation and safety: A foundation for safety and vehicle-infrastructure integration research. In *Intelligent Transportation Systems Conference, 2006*. IEEE, 9–16.

[13] M.C. Nussbaum. 2001. *Women and Human Development: The Capabilities Approach*. Cambridge University Press, Cambridge.

[14] Claudia Pagliari, Don Detmer, and Peter Singleton. 2007. Potential of electronic personal health records. *BMJ: British Medical Journal* 335, 7615 (2007), 330.

[15] Stuart J. Russell. 2017. Provably Beneficial Artificial Intelligence. *Exponential Life, The Next Step* (2017).

[16] A. Sen. 1999. *Development as Freedom*. Oxford University Press, Oxford.

[17] Richard H Thaler, Cass R Sunstein, and John P Balz. 2014. Choice architecture. (2014).

[18] The IEEE Global Initiative for Ethical Considerations in Artificial Intelligence and Autonomous Systems. 2016. Ethically Aligned Design: A Vision For Prioritizing Wellbeing With Artificial Intelligence And Autonomous Systems, Version 1. IEEE, http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html.

[19] Ibo van de Poel. 2013. *Translating Values into Design Requirements*. Springer Netherlands, Dordrecht, 253–266. https://doi.org/10.1007/978-94-007-7762-0_20

[20] Jeroen van der Hoven and Noemi Manders-Huits. 2009. *Value-Sensitive Design*. Wiley Online Library.

[21] Wendall Wallach and Collin Allen. 2008. *Moral machines: Teaching robots right from wrong*. Oxford University Press.

[22] Michael Winikoff. 2017. Towards Trusting Autonomous Systems. In *Fifth Workshop on Engineering Multi-Agent Systems (EMAS)*.