Contents lists available at ScienceDirect

Technology in Society

journal homepage: http://www.elsevier.com/locate/techsoc

Developing a roadmap for the moral programming of smart technology

Bart Wernaart

Fontys University of Applied Sciences, Eindhoven, the Netherlands

ARTICLE INFO

Keywords:
Moral programming
Ethical decision making
Moral intensity
Smart technology
Normative ethics
Meta ethics

ABSTRACT

Smart technology is increasingly integrated in our ethical decision making. This raises questions as to how we should morally program technology. Deciding on moral programming depends on the moral intensity of the ethical issue. A moral intensity dashboard for engineers can help allocate the most suitable moral authority for a particular moral programming. Technology is not capable of 'doing' ethics the way humans do. This leaves forms of consequentialism and deontology as the most reasonable programming alternatives, using deontic logic as a starting point. Furthermore, it is very likely that in the more complicated settings, technology should have elements of meta ethics in its moral programming to adequately deal with scenarios that lead to conflicts in moral programming. We propose to use the calculation methods that stem from a comparative approach or the Expected Moral Value approach. All this has considerable consequences in how we should see moral programming in technology-driven ethical decision-making processes. We will therefore propose a roadmap for the moral programming of smart technology.

1. Introduction

Our society has transformed from an information society into a smart society [1]. Within a smart society, technology and humans are constantly and structurally connected. This connectedness appears in every aspect of life. After all, it isn't just technology that is changing rapidly; it is the entire sociological system that is changing [2]. Artificial Intelligence, amongst others, operates as a catalyst for this change within all the domains in which people and smart technologies are interconnected [3]. This leads to a change in moral decision-making in which not only humans but also technology can be part of ethical decision-making (EDM) processes [4]. This results in new ethical challenges: due to the ever-increasing complexity of smart technology, human interacting with this technology is not always aware of its nature (human or machine), as well as the moral authority and moral foundation of the ethics used in smart technology.

The idea of smart technology that is able to make ethical decisions amazes and also frightens people [5]. In academic literature, the discourse on moral programming mostly focusses on laboratory-type, all-or-nothing and life-or-death situations in which mostly a programmable choice between consequences or principles must be made (see for a notable example: [6]. The focus is on grandiose themes in which people's lives are seriously affected by the ethical choices made by machines [7]. Examples are the warbot that has to make a decision when to shoot and who to shoot in a warzone [8], or the self-driving vehicle, that has to decide what to do in case of an imminent crash. Such examples ignite the imagination, and due to their extreme outcomes are a fertile ground for a sharp ethical discourse. However, moral programming mostly results in more subtle consequences that are not a matter of 'life or death' in nature. Current academic discourse on the more practical complications of moral programming is scattered and mostly focusses on particular technological products that are used in various applied contexts. Their moral programming may have serious consequences for the user and/or society, such as the a self-driving vehicle discussed above [9], the sexbot (see for a discourse: [10], the carebot [11], or the domestic robot [12]. Furthermore, we see a rather scattered debate about particular issues that may relate to the consequences of moral programming. These include privacy issues concerning the usage of social media or the adoption of technology [13], or the impact on the consumer's autonomy and privacy by marketing strategies using a high level of automated processes, including big data [14], the ethics of gamification [15], or the ethics of using techniques to replace human workers [16]. These debates have one thing in common: they focus on the morality of societal impact caused by smart technology, but not so much on how the ethics were programmed in the system causing these consequences.

Moral programming implies an ethical decision-making (EDM) process in which smart technology is used in ethical decision making. There

E-mail address: b.wernaart@fontys.nl.

https://doi.org/10.1016/j.techsoc.2020.101466
Received 30 June 2020; Received in revised form 16 September 2020; Accepted 25 November 2020
Available online 9 December 2020
0160-791X/© 2020 The Author(s). Published by Elsevier Ltd. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

is an abundance of literature that discusses EDM models in the context of human ethical decision making, especially in the context of business [17–20]. In these models, various stages of ethical decision making are proposed, as well as which factors influence the EDM process. This has led to a significant amount of research in which researchers attempted to validate these factors [21–24]. These EDM models are not easily applied in the context of moral programming, since they are about individual, human, decision-making, and not automated, programmed decision processes.

Therefore, what is lacking so far is a more coherent discourse on ethics regarding the moral programming of smart technology that focusses on the more practical, everyday type of cases, and not so much the laboratory type, life-or-death situations. After all, we assume that the average individual will increasingly have to deal with for instance automated B2C processes in which ethics is programmed than being confronted with a warbot or a self-driving vehicle that has to choose between killing or injuring. Our goal is to develop a roadmap for the moral programming of smart technology. Smart technologies exist in many ways and forms. In this contribution, we focus on smart technology that is able to analyse and monitor a situation and interact with humans. We use the chatbot as a continuous example. The chatbot is one of the smart technologies that drives on artificial intelligence with which the individual is literally confronted the most [25]. In our journey towards a roadmap, we discuss several theoretical aspects regarding the 'when?', 'who?', 'what?' and 'how?' with respect to moral programming. To explain these theoretical aspects, we will talk about the fictive chatbot 'Sylvia'. She is able to respond to questions of consumers and advise them on the products they can purchase.

2. Moral problems

A first matter we need to discuss is *when* we consider the programming of smart technology to be *moral* programming. For this, we need to distinguish between moral programming and the consequences of programming that may cause ethical issues. Using a carebot to help elderly people may cause a moral problem. For instance, it might 'dehumanize' the healthcare profession, and increase a sense of loneliness amongst the elderly when they are increasingly nursed by robots instead of humans [11]. These social consequences can undoubtedly lead to a moral debate. However, the value judgement that needs to be made here - e.g. what costs should society bear in order to guarantee sufficient human aspects in healthcare for elderly people?- is, in the end, done by human actors, and not by machines. In other words: the use of smart technology leads to a moral problem, but the technology is not per se programmed to solve this problem. In this contribution, we consider programming to be *moral* programming when technology is programmed to (partially) solve a moral problem.

A second issue we need to discuss here is how we classify a moral problem. In academics, there has always been an intense debate regarding the existence [26] and nature of a moral problem [27]. As a result of this discourse, a variety of terms is used to address the situation in which a person is unsure which moral action to pursue, such as a moral dilemma, a tough case, or a moral conflict. The existence and nature of such moral problems greatly depend on the level of absolutism someone accepts in normative ethics, and the level of comparability one assumes when more than one ethical theory offer solutions to a moral problem. When we assume that moral facts exist, and there is absolute truth to be found in morality, a moral problem cannot exist [28], and when we assume that moral theories are incomparable, a moral problem cannot be solved [29,30]. In this contribution, we prefer to use the more general term 'moral problem'. We propose that –at least in the context of programmed ethics-moral problems do exist and can be solved, as will be discussed further below (in particular in section 4.2. in the context of normative ethics and 4.3 regarding theories in meta-ethics).

2.1. A moral judgement

A moral problem requires moral judgment in order to be solved. However, it becomes unclear what the required action should be in a given moral problem when there seems to be more than one available alternative, when all alternatives are morally right considered in light of the moral theory that proposes the solution, and only one action can be executed.

Let us turn to our fictive chatbot Sylvia. A great deal of the communication between Sylvia and the customer she interacts with will be more about factual issues and not so much about value judgments. For instance, when a customer asks about the size of a bookshelf, the chatbot will use the factual information that is available and communicate this accordingly. We can hardly say there is ethics involved here. From a moral perspective, things become challenging when Sylvia takes a decision in her relationship with the consumer that is clearly beyond facts. Imagine Sylvia is a chatbot used by a company that sells infant nutrition: the fictive chatbot Sylvia is designed to advise mothers on how to feed their infants when natural breastmilk needs to be complemented or replaced by infant formula, using all the available (scientific) knowledge in the field of infant nutrition. A mother could ask Sylvia for advice regarding the dilemma she is facing: her breasts produce too much milk so that the child almost chokes on the milk, pumping the milk and then feeding the child through a bottle is time-consuming and results in a lack of sleep and a lot of stress, and infant formula solves the time and stress problem but is proven to be not as healthy for the child as breast milk. Now, imagine the company to which Sylvia 'belongs' wants their chatbot to give the best possible advice that contributes to building a reliable image of the company. Sylvia could, for example, give the advice to continue the breastfeeding, since this is simply the best nutrition for the child with long-term effects for its health. At the same time, it could, however, lead to significant stress and difficulties in a healthy attachment between mother and child, which may also have negative long-term consequences that cannot be predicted precisely. On the other hand, she could advise to (partly) use the infant formula offered by Sylvia's company, resulting in a less healthy nutrition pattern for the child, but at the same time encouraging a less stressful home situation. This also may have long-term effects that are unclear at this moment. In general we could say that the advice involves a value judgement, and both sketched alternatives can be defended using a theory in normative ethics (also named 'moral theory'). For instance, we could say that from a utilitarian perspective, Sylvia should advise to do what leads to the greatest happiness for the greatest number. Obviously, the mother will be relieved with the reduction of stress in her home situation, and as for the child: infant formula is not the best, but most certainly the second-best nutrition. Advising to continue beast-feeding will only result in a relatively small increase of happiness for the child in the long term, while it will result in a large decrease of happiness for the mother. However, when we use a deontological approach, say a right based approach, we could argue that the child has a fundamental right to breast feeding since this is the best possible nutrition for the child, and the mother has a fundamental right to give mother milk, which should be encouraged and facilitated where possible. This is a fundamental principle, and no matter what the consequences are, should be respected.

In the case of the infant formula, there are at least two alternatives – one is based on a utilitarian approach and the other on a deontological approach – and at first glance it seems impossible to neutrally consider that one option is better (or worse) than the other. This leads to three urgent questions in the moral programming of smart technology that need to be addressed:

1. *Who* has the moral authority to solve a moral problem through moral programming? This matter is further discussed in section 3.
2. *What* can be decided when we need to solve a moral problem through moral programming, and

3. *How* can the moral judgment be programmed in the smart technology? These matters are further discussed in section 4.

3. Moral authority

The way smart technology is morally programmed has consequences for various stakeholders. This includes the engineer that programs the technology, the company that makes use of the programmed technology, the consumer that interacts with the programmed technology, and possibly society in general, that bears some of the consequences of how the machines are programmed. The question is: *who* should eventually decide on moral programming? Millar [31] noted that while engineers are the ones who will have to carry out the programming and have the expertise to do so, they do not necessarily have the moral authority to decide how to program ethically. In other words: technological expertise or capacity to morally program technology is not the same as moral authority (also [32]. Within this context, the subdivision Millar makes between high stake and low stake ethical settings is useful (see Fig. 1). This subdivision can be further nuanced using the concept of 'moral intensity' as introduced by Jones [18] in EDM-theory.

3.1. Low-stake settings

In case of the low-stake settings, those who are affected by the ethical decision are relatively indifferent regarding the moral choices made in the programming because it does not (or barely) affect values that are important to them. Millar [31] proposes that in the case of low-stake ethical settings, the engineer can decide on moral programming. This is by far the most practical approach and has very limited ethical consequences for the involved stakeholders.

Let's go back to our chatbot Sylvia, who now operates as a shopping assistant and gives advice on sound systems. An example of a low-stake ethical setting could be the moral question whether or not a chatbot should take the energy consumption of the various sound systems into consideration. What to do if sound system A has a slightly higher energy consumption but a significantly better price-quality balance compared to sound system B, which is mildly lower in energy consumption, but also has a lower price/quality balance? While energy consumption could lead to extra pollution, the differences in pollution are relatively small, and also depend on the consumer's type of energy supply. The moral problem here could be whether negative environmental impact should be balanced with consumer comfort. Which of the two would lead to the greatest happiness of the greatest number, and how could we measure that? Whatever the moral judgments will be, the stakes are relatively low for all the stakeholders and therefore -at first glance-the programming can easily be entrusted to the engineer. Most probably, the engineer works in a business environment and has to act in the interest of the organization he works for. This means that in practice, when deciding on how to program, values such as prosperity (profit) or (customer) satisfaction will play an important role.

3.2. High-stake settings

One could say that it is acceptable when business interest is used as a driver for ethical decision making in low-stake ethical dilemmas, while it is less acceptable when this happens in high-stake ethical decisions. In case of high-stake values, the core values of those who are involved are affected, and stakeholders are not as indifferent to the consequences of the ethical decision that can be made compared to low-stake settings. In such cases, the engineer may subject users to undesirable paternalistic relationships by imposing personal ethical views on its users [33]. This can be particularly problematic considering that engineers usually represent a rather homogenous group [34], which could result in moral programming that is laden with stereotypical concepts of –for instance-gender [35] or race [36]. Therefore, if ethical decision making in high-stake settings is part of the design of technology, those who are affected should be involved in the decision-making process [37]; in the particular context of health robots, see Ref. [38]. Millar [31] argues that the moral authority to take a decision on moral programming depends on who is affected. If the effect predominantly relates to the individual consumer that interacts with the technology, this consumer should be enabled to influence the moral programming in such a way that the norms and values of the consumer are respected through the programming. This could (for instance) be realized by offering various user-preferences that can be selected by the consumer, each representing a different emphasis in ethics. An example would be the various privacy settings a user can opt for in smart devices, each representing a different balancing of privacy with other values, defining the decisions the device will take regarding the use and safety of certain types of personal data.

Things are different when the consequences of the moral decision are more societal. In those cases, moral programming needs political recognition through a democratic process. The authority of the individual consumer ends and the work of the law maker begins. For instance, what would happen if Sylvia is programmed in such a way that she could lie, or leave out crucial information in a conversation with consumers, in order to make them do the right thing [39]? In the case of the sound system chatbot, would it be ethical to leave out positive quality aspects of products that are less sustainable, so that people tend to choose a more environmentally friendly device? In other words: is the programming of a 'white lie' to contribute to morality an acceptable means to this end? When chatbots, or smart technology in general, can be programmed to lie in order to realize what is right according to their own programming, the resulting manipulation of the consumer may be unlimited. This moves beyond the sphere of the individual, since we can all be affected by lying technology, assuming that in the near future we interact daily with smart technology and moral programming. Because the consequences affect society in general, it is only reasonable to

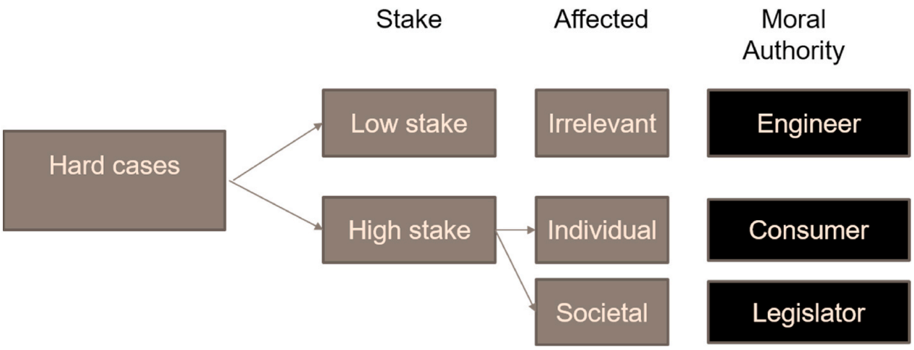

Fig. 1. The moral authority of the moral stakeholders in moral programming.

assume that the lawmaker of that society decides on the permissibility of lying machines.

In practice, there could be a thin line between high- and low-state cases, and they will have to be defined carefully per sector, product or service, especially when moral authority is claimed by more than one stakeholder, or when an individual case may be qualified as a low-stake setting, while the total sum up of a lot of low-stake settings may result in a high-stake situation (perhaps the question whether negative environmental impact should be balanced with consumer comfort is an example of the latter). Therefore, a more nuanced approach could be required, which can be offered by exploring the concept of 'moral intensity' as introduced by Jones [18].

3.3. *Towards a moral intensity dashboard for engineers*

Jones [18] was amongst the first to recognize that the characteristics of the moral issue play a crucial role in ethical decision making, besides the already widely explored traits of the moral actor and organizational factors.

3.3.1. *EDM-theory*

In EDM theory, the ethical decision of individuals is analysed. Naturally, there is a strong focus on the particular traits of that individual. In literature we see a diverse terminology to describe these traits, such as 'moral capacity' (subdivided in 'moral character' and 'integrity capacity') [17,40], 'moral sensitivity' [41] or 'moral imagination' [42]. What they have in common is that they relate to individual factors that influence a rational or intuitive/emotional ethical decision-making process [17,43]. Such individual factors have been defined numerous times in literature, and are mostly a mixture of demographics, psychological factors and ethical experience (for a more detailed list: [17,21]. Next to individual factors, organizational factors that influence an individual's ethical decision process are defined in literature [20,23,44,45], such as a reward system, work roles and organizational culture [46]. Such individual and organizational factors are extremely relevant when considering the individual ethical decision-making process of an engineer. However, in this contribution we do not try to explain individual ethical decision making but instead seek to offer a structuralized roadmap to allocate moral authority in the context of moral programming. For that, we assume that issue-related characteristics can offer us more nuanced criteria to assess the stake of a moral problem. Jones' concept of moral intensity identifies the effect of issue-related aspects on decision making. The hypothesis is that when these aspects are more intense, ethical decision making is more urgent. In other words: the stakes are higher. This hypothesis is empirically tested by various authors, and while their conclusion is that not all dimensions equally effect the ethical decision making, they are all considered to be influential (see for instance the findings of [47] or [48]. Where Jones focusses on effect, we focus on characterizing the moral problem in itself in order to conclude who the moral authority in a given moral programming context should be.

3.3.2. *Six dimensions in moral intensity*

According to Jones [18]; there are at least six dimension that altogether define the moral intensity of a situation.

1. Magnitude of Consequences: what is the total sum of harms or benefits for the moral stakeholders? In the case of Sylvia advising on a sound system, we considered the moral dilemma to what extent the negative environmental impact should be balanced with consumer comfort in a low-stake setting. When we consider the overall impact of the entire product however, (and not just one sales unit) or perhaps even the environmental impact of the branch as such, the magnitude of the consequences is significantly higher, and we might reconsider labelling the dilemma as a low stake setting.
2. Social Consensus: what is the degree of social consensus of a particular moral action? In the breastmilk case, there is a societal consensus that breastfeeding is extremely important for the healthy development of the child. It is even considered to be a human right to give and receive breastmilk [49], meaning that the mother should have the freedom to do so, and make her own informed choice on the matter, without too much interference of marketers trying to sell substitutes to breastmilk [50].
3. Probability of Effect: what is the probability that the effect will indeed take place as a result of the moral action? While it is known that breastfeeding is the best nutrition choice for the child, it is uncertain whether a child that was raised with infant formula instead of breastmilk will indeed develop in a less healthy way [51].
4. Temporal Immediacy: what is the timespan between the moral action and the moral consequences? This dimension relates to the sense of urgency in the moral decision making. In the case of the sound system, the consequences of CO2 pollution are not the day after the sales but cover a longer period of time. The same can be said about the possible health effects of the nutrition choice for the child however; the nutrition choice has an immediate effect on the stressed-out mother. While this temporary effect can be used to explain that people would generally feel more responsible in situations that involve immediate consequences, long-term consequences can be just as (un)ethical as short-term consequences, which says very little about whether an ethical dilemma involves a high or low stake setting. However, this dimension generally expresses a degree of urgency of the moral problem. We therefore propose, in line with Mitchell et al. [52] -who also identified the dimension 'urgency' to express the need for immediate action in the field of business ethics and stakeholder analysis-to broaden the scope of this dimension to 'degree of urgency'.
5. Proximity: what is the sense of nearness towards the moral stakeholders that will be affected by the moral action? This dimension is – by far - the most subjective one. The main idea is that in ethical decision making, moral actors usually feel more responsible for moral stakeholders who are closer to them. In this case, the intended moral actor is a machine who has no initial preferences or a sense of nearness towards moral stakeholders, and is programmed by an engineer, who –on the contrary-may have such preferences. In our view, this dimension is less useful in classifying the stake of the moral issue, and instead could better be classified as an individual or organizational factor: individual, when personal convictions or relations result in a higher degree of proximity, organizational when the business environment has this effect.
6. Concentration of Effect: what is the level of concentration of the effect of the moral action? In the sound system case, the CO2 pollution affects a large group of people a little bit, while in the breastfeeding case, a small group (especially the mother and the child) are deeply affected by the choices made. Jones added this dimension assuming that moral actors generally feel more responsible for their actions when a certain magnitude of the effect is highly concentrated on a small group (or even one individual) compared to when the same magnitude of effect has a low concentration, simultaneously affecting a large group of people who individually barely notice the consequences. For our purpose, this dimension can be useful to assess whether the effects are predominantly individual or societal.

In essence, Jones focuses on three main issues: the consequences or effect of the moral action (dimensions 1, 3, 4, 6), the social perception of morality (2) and the sense of nearness towards the moral stakeholders (5). At first glance, we could say that there is a strong emphasis on the consequential aspects of a moral action [17], while in ethical decision making we would expect that more deontological elements are also situational factors that may play a role, such as human rights and equality [53–55]. Such deontological aspects are, however, represented in the degree of societal consensus towards a certain moral issue and are therefore implicitly present in the model of Jones [56]. Another issue is that there is a certain overlap between issue-related factors, organizational factors and individual factors, and most models do not draw a clear line between them (see for instance: [45] considering the differences between organizational and issue-related factors). As we saw in the model of Jones, individual perception cannot be isolated from issue-related factors. This is especially notable when considering the 'proximity' dimension, but also the intensity of the other dimensions depends on one's individual perception. Therefore, moral intensity is almost per definition a perceived intensity, depending on the individual who assesses the moral issue [57]. It seems unlikely - and undesirable - that we can neutrally characterize moral issues as high or low stake without the subjective influence of the individual making a moral decision and isolate the ethical issue as a separately existing phenomenon. However, Harrington [58] found that in cases when there is a strong societal consensus or when the seriousness of the consequences is high, individual features of the decision-maker become less relevant. A high stake setting therefore is expected to be more unanimously recognized, regardless the individual backgrounds of the engineer.

We can try to help the individual engineer who programs the moral actor - the smart technology - to be aware of, and assess in more detail, the moral stake that corresponds to a moral issue, and consequentially understands who ideally should be the moral authority that decides on the nature of moral programming. To this end, we drafted a 'moral intensity dashboard for engineers' (Fig. 2), which could be a starting point for assessing the stakes involved in moral programming more structurally and allocating a moral authority to particular moral programming issues. In essence, this dashboard could help categorize a moral issue as 'high stake' or 'low stake' by assessing the dimensions 'society', 'effect', 'probability' and 'urgency', and specify whether the consequences of the moral programming are predominantly individual or societal by assessing the 'concentration of effect' dimension.

4. Moral theory

The next question that needs to be answered concerning the business of moral programming is *how* we can choose a moral theory that should be used in the programming of smart technology. In normative ethics, ethicists propose an answer to the question 'what is the right thing to do?'. Various theories in normative ethics propose an answer to this question from different and opposing perspectives. Universalist and absolutist theories can be applied consistently regardless of who uses them, while the more relativist theories recognize that 'the right thing' depends on the particular traits of the moral agent. To answer the question how we can choose a moral theory in moral programming, we first need to explore the nature of normative ethics and how this relates to the relation between humans and machines.

4.1. *Universalism, absolutism and relativism*

While there are many convincing approaches in normative ethics, there is a strong emphasis on universalist and absolutist theories in Western academic literature [46,54]. These theories find their origin in enlightenment thinking and offer fundamental moral rules in life that can be applied by every individual in any given situation (universalist theories) and/or embody an absolute moral truth (absolutist theories) [59]. Consequentialist theories, such as utilitarianism and egoism are typically universalist: something is morally justified when it leads to the best result for the moral actor (egoism) or the greatest number of people (utilitarianism). While these theories do not encompass an ultimate moral truth, they can be applied the same way in any given moral dilemma.

Deontological theories, such as duty ethics, moral rights and the principle of equality, could be considered both universalist and absolutist: they claim to hold moral truth, and can be applied in any given moral dilemma. The well-known categorical imperative, introduced by Immanuel Kant, is an interesting example of duty ethics. The main idea is that human beings have a duty to only act according to that maxim whereby you can, at the same time, will that it should become a universal law [60].

Deontological ethics in which moral rights are proposed as a yardstick for moral behaviour finds its origin in the 'social contract' as introduced by Thomas Hobbes, Charles de Montesquieu and Jean Jacques Rousseau: people should give up a part of their sovereignty in exchange for their protection of fundamental rights. These rights are mostly referred to as human rights, and stern from the fact that people are human beings and for that reason have such rights [61]. All moral actions should lead to the respect of these rights and contribute to its fulfilment. Deontological ethics based on the principle of equality was introduced into most detail by Rawls [62,63] and assumes that various interpretations of equality should lay at the core of moral actions - mostly at societal level. In general, deontological theories can be used in any given moral dilemma and hold an absolute moral truth that excludes other theories on normative ethics.

Not all theories on normative ethics are universalist or absolutist. Some of them have a more relativist nature [64], accepting that there is no moral truth or universal application of morality, and morality therefore depends on the individual or society.

For example, virtue-based ethics [65,66], post-modernist ethics [67,68], relation ethics [69,70] and -to a certain extent-discourse ethics [71] have characteristics of moral relativism [59]. However, it needs to be

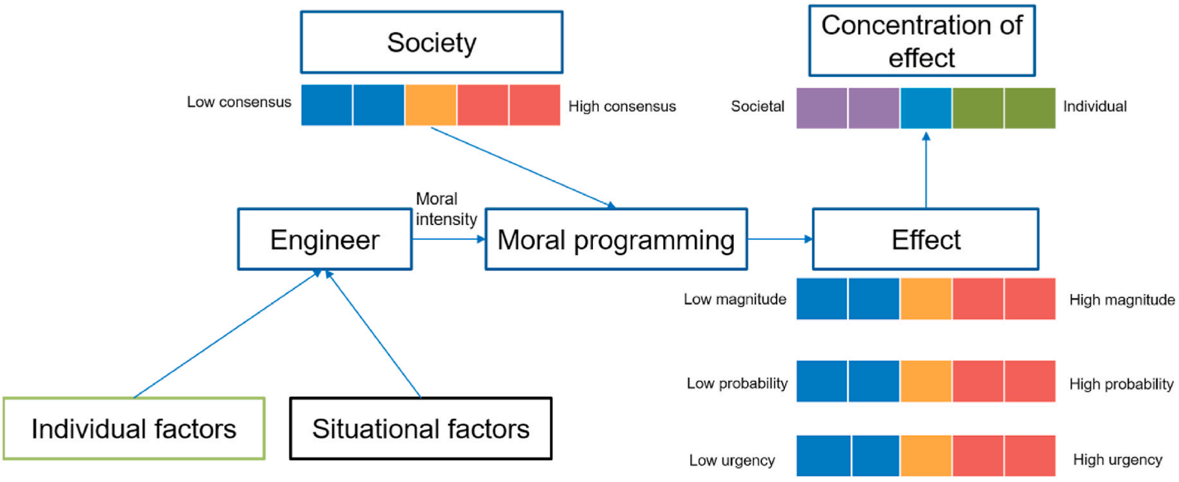

Fig. 2. Moral intensity dashboard for engineers.

noted here that labelling an approach in normative ethics as exclusively relativist leads to an overly simplified debate that does not do justice to the particularities of each approach. As we will see below, some approaches in normative ethics - most notably virtue ethics [72,73] and discourse ethics [74]- have absolutist or universalist components or try to overcome a bridge between relativist and non-relativist theories. What these theories have in common however is that they at some point answer to critiques of theories in normative ethics that where a product of enlightenment-thinking, mostly for being too inflexible and consequentially ignoring individual or cultural factors that may affect the notion of what is right or wrong. Furthermore, these theories we label here as being more relativist, somehow assume that human characteristics or relations are at the core of establishing morality: a notion that is of particular importance in the context of this contribution.

4.2. *Human versus machine*

The idea that human beings who take ethical decisions are mainly driven by universalist and absolutist theories in normative ethics is increasingly questioned. At the same time, it seems that those exact theories are suitable for moral programming.

4.2.1. *Humans and normative ethics*

Using arguments from cognitive science, Johnson [75] said that it is very unnatural and unlikely that moral behaviour of human beings is driven by universalist or absolutist theories. He observes that such theories usually work in simplified 'laboratory' settings. An example of such a setting is the famous Trolley Problem [76], in which the moral actor has to choose between not pulling the switch - knowing that five people will be overrun by a train - or pulling the switch - saving these five but sacrificing the life of one person who is on the other track. This dilemma has been altered many times in moral philosophy but is mostly exclusively used to demonstrate the differences between consequentialist and deontological normative ethics. Currently, we see variations to this theme in the academic literature on ethics and technology, for instance in the context of ethics and self-driving vehicles [7]. According to Johnsen, such experiments where designed to make the universal theories work, but by no means represent a real-life situation. Instead, he argues that human beings make moral decisions based on their moral imagination, which is most certainly not a fixed thing, since *'Human beings are not fixed quasi-objects that have an independent prior identity and then go about making choices from which they are distanced. We are, rather, beings in the process whose identity emerges and is continually transformed in an ongoing process of reflection and action'* [75]; p.148). Johnson argues that even though universal and absolutist theories can be a source of inspiration, or even influential in the development of someone's moral imagination, they are at most expressions of shared moral assumptions of Western enlightenment thinking. To assume that they are therefore universal would be erroneous reasoning: absolute moral theories could at most be the result of shared cultural values, not the origin of them. Instead, he argues that prototype structures of concepts we recognize in our brains as 'typical situations' that evoke a certain emotion and require a certain moral solution. These are very personal, and mostly determined by the life experience you had so far. Such prototypes are put in a certain context, depending on what you know about the situation, which leads to semantic frames based on which we perceive the situation. We then try to learn from these experiences and draw lessons from it, which we may apply in the next situation - or not, depending on the situation. A moral theory should not answer the question 'what is the right?' but rather offer a framework to recognize, structuralize and expand one's moral imagination. Therefore, Johnson argues that most moral decisions are made without universal rules; they originate from the moral imagination of the involved actor.

This view – which we already touched upon - is supported by most theories on EDM processes. There is an abundance of literature in which EDM procedures are described, and most of them take Rest [19] as a starting point, where the process of ethical decisions is composed of four steps: 1) recognizing the moral issue, 2) making a moral judgment, 3) establishing moral intent, and 4) moral acting. While there are some variations to this theme (for an overview, see: [18] the actual discourse is not so much about the stages of EDM processes, but rather concerns the identification of factors that influence these stages. Factors that influence individual ethical decision making are hardly coherent theories in normative ethics. While such theories are credited to be used as a reflective tool to validate or justify an ethical decision [43], they are not at the core of the EDM process itself. Instead, we see an emphasis on various factors that influence EDM processes that suggest a very flexible, case by case, approach towards ethical decision making, fully depending on the individual features of the moral actor, its environment and the particularities of the ethical situation [17]. Therefore, we could argue that the answer to the question 'what is the right thing to do' does not stern from a coherent and rigid theory in normative ethics, but depends on coincidental particularities: it is of a relativist nature.

4.2.2. *Smart technology and normative ethics*

In a sharp contrast, most literature on moral programming focusses on the universalist and absolutist theories, mostly underlining the different outcomes of consequentialist and deontological programming. After all, theories in normative ethics that are universal and can be imposed on someone (or something) are more likely to be suitable for programming compared to normative ethics that originate from the moral imagination of human beings. For instance, there is a vivid debate about the ethical programming of self-driving vehicles [77], in which mostly ethical dilemmas are explored that roughly relate to the aforementioned Trolley Problem: should the car in case of an imminent crash choose for the outcome that leads to the greatest happiness of the greatest number, or should the car respect certain universal principles? While we could argue that Trolley Problem scenarios are by no means a reflection of real-life situations [9], the debate embodies a good starting point on how to program the self-driving car when it has to make a moral judgment.

To translate absolutist normative ethics to programming language, an often proposed approach is to make use of deontic logic (a term introduced by Von Wright, 1951). This means that ethical programming is narrowed down to four main categories: obligatory actions (o), permitted actions (p), forbidden actions (f) and actions that are morally indifferent (i) [78]. For instance, Powers [79] proposes to use deontic logic to create so called 'Kantian machines' that are able to apply - with some modifications - the categorical imperative when they function in any given situation. This would mean that our chatbot Sylvia is programmed in such a way that it recognizes actions that contradict the categorical imperative and classifies them as forbidden. The lying machine would be a good starting point. If Sylvia could lie, and we would quantify lying in an individual case according to universal proportions, it would lead to a world in which ethical machines become unreliable, which would interfere with the reason of existence of the machine itself: human-built machines relying on their functionality. Considered in this quantified proportion, Sylvia would destroy her own functionality by lying, which would be pointless. The engineer should therefore be able to program the machine in such a way that it would refrain from doing things that contravene its functionality. The opposite, telling the truth, is then automatically an action that is obligatory - since the negation of telling the truth would constitute a lie. If communicating data to a consumer does not interfere with telling the truth, and does not constitute a lie, it is permitted or morally indifferent. This means that for instance framing the truth in an attractive way would be permissible, as long as it does not lead to a forbidden action. In deontic logic there is a certain overlap between the categories 'morally indifferent' and 'permitted' and 'obligatory': all morally indifferent action are permitted, but not vice versa, since obligatory actions are also permitted actions but not morally indifferent [80], as we can see in Fig. 3.

The use of deontic logic as a means to create ethical machines has its

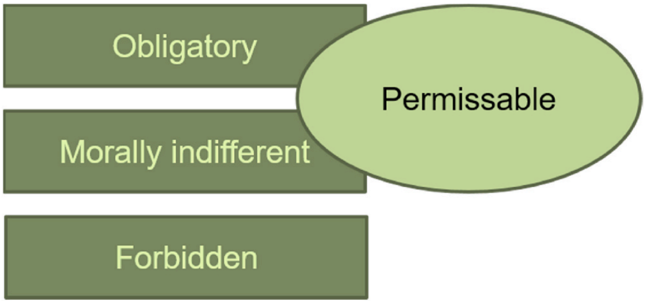

Fig. 3. Overlapping categories in deontic logic according to Von Wright [80].

limits. Bringsjord [81] argued that while we could use elements of normative ethics in programming machines, this will only result in relatively 'flat' moral actors. One thing a machine that is programmed with deontic logic will never do is something that is in itself morally right, but not wrong not to do. Such actions are mostly unaccounted for in the absolutist theories [82], and more in particular in deontic logic. However, they are considered to be morally praiseworthy to do [83]. In fact, such actions could be labelled as supererogatory actions, and cannot be programmed using deontic logic. After all, they are permissible to do, but not morally indifferent nor obligatory.

The main reason for this is that a machine as a conscious moral actor does not have the same consciousness compared to a human being. Ned Block [84] observes that consciousness can be subdivided in so called access conscious (A-conscious), and Phenomenal-consciousness (P-conscious). The first involves an awareness that is required to reason and rationally guide speech and action. Or as Bringsjord [85] would put it: a 'zombie state of mind'. To a certain extent, we could program machines with an A-conscious. In contrast, P-consciousness is the capacity to experience things and have subjective awareness. This is what makes us human and is something a machine in the foreseeable future can never have [85]. An example: an A-conscious mind could recognize the colour red, and act accordingly (for instance, stop the car), and communicate its actions. A P-conscious mind is able to experience what it feels like to observe the colour red. We may be able to create machines that imitate the consequences of experiencing the colour red, but it seems unlikely that this experience in itself can be - directly or indirectly through self-learning - programmed. To rephrase that in ethical terminology: a machine may act in a moral way but cannot be moral in itself [86]. For being moral, a P-conscious mind is required. It is exactly the P-consciousness that forms the basis of theories in normative ethics that are of a relativist nature, since life experience is one of the core elements of relativist theories (most notable in post-modernist ethics).

In conclusion, see also Fig. 4, we can state that 1) machine ethics is most likely based on different normative ethics when compared to human ethics (i.e. absolutist theories versus non-absolutist theories), and 2) machine ethics is relatively 'flat' compared to human ethics, making smart technology, also self-learning technology, unable to escape the patterns of deontic logic.

Since we are living in an era in which machines *replace* human beings, this is per definition inaccurate in the field of ethics: human ethics cannot be replaced by machine ethics. When a job previously done by a human being that involves the interaction with consumers and in this interaction human ethics is applied is now done by a chatbot, human ethics is *removed*, and machine ethics is *added*. From the above, we can only deduce that it is reasonable that when smart technology interacts with humans in the context of a moral problem, humans should have the right to know they are interacting with a machine, and not another human being.

4.3. How do we choose which approach in normative ethics should prevail?

In the previous section we discussed which approaches in normative ethics are most likely to be eligible for ethical programming. However, the eligible moral theories may lead to opposing conclusions when applied to a moral problem. We have seen the example of the chatbot advising whether or not to use infant formula as an alternative to breastmilk. We explored two moral approaches: a utilitarian -consequentialist- approach in which one could argue that the chatbot will advise the mother to replace (part of) the breastmilk by infant formula, and a right-based –deontological- approach in which the continuation of feeding breastmilk would be advised. In such a situation, we cannot possibly program the chatbot in such a way that it complies with both approaches. This means that a choice must be made. Hypothetically, the relevant moral actor could decide to either go for the utilitarian or right-based approach, based on their ethical preferences. However, it is nearly impossible to program such choices per potential future situation or conversation. On the other hand, programming a machine with only one approach in normative ethics can also lead to undesirable results, since different situations may require a different approach in ethics. This leads to the question how fixed or flexible moral programming should be [31]. In other words: do we make a fixed choice in programming normative ethics, with permanent principles and/or end results, or do we allow the machinery to reflect on its own programming, within the boundaries set by meta-ethical programming?

4.3.1. Meta-ethical programming

In case of a moral programming, the choice for an approach in normative ethics is an arbitrary one, depending on the preferences of the moral actor who has the moral authority to make a decision on the matter. Especially in the more complicated ethical machines, we can

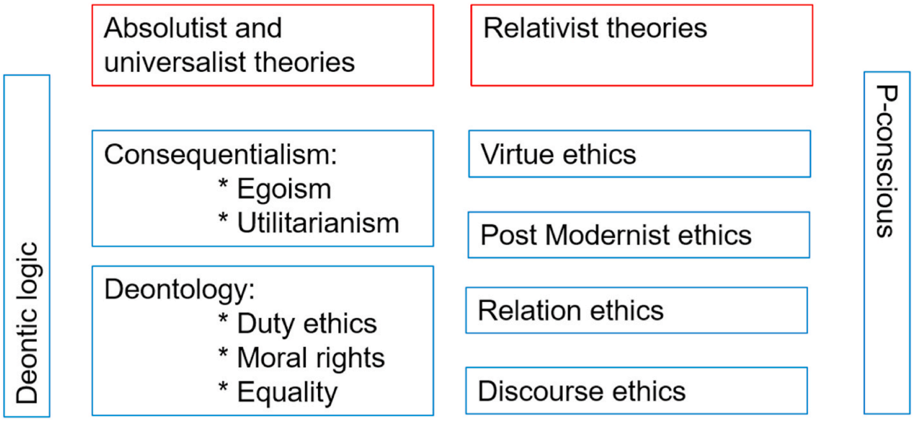

Fig. 4. Normative ethics and moral programming.

assume that it makes sense to include meta ethics in the programming of technology [87]. After all, the deontic logic that is used to program the bot can lead to conflicting outcomes. Consider the situation that our chatbot Sylvia is programmed with two obligations:

1. Never give advice on something that violates the right to health or healthcare, in which part of this right is understood as the right for breastfeeding or receive breastmilk.
2. Advice should lead to optimizing consumer satisfaction.

In the case of the worried mother described in section 1.2., it is almost impossible to fulfil both requirements, for the consequentialist aspect (2) conflicts with the deontological one (1) in almost any advice that can be given: advising on using infant formula violates the deontological obligation, where advising on continuing breastfeeding will continue the stressful family situation, and not lead to a happy customer. This means that at some point, it would be helpful if the chatbot would be able to reflect on its own moral programming: this means that there should be some meta-ethical programming.

This leads to the question whether it is possible and desirable to program a machine in such a way that it is able to compare different ethical solutions to a moral problem, and evaluate which solution is most suitable for the case at hand (flexible programming). This question touches upon a discourse in meta ethics: can we compare what is right or wrong according to one moral theory with what is right or wrong according to another? For instance, can we say that a consequentialist approach is better, worse or equally right compared to another approach in normative ethics [29,88]? This problem has had different names in literature so far, but they all address more or less the same issue. Examples are 'the problem of moral uncertainty' [87], 'value incommensurability' [88], or 'the Problem of Inter-theoretic Value Comparison' [89]. For a long time, the leading view was that we cannot possibly compare different approaches in normative ethics, and that moral action exclusively depends on which approach in normative ethics is preferred by the moral actor. In practice, this means Sylvia would simply have to choose between consequentialism and deontology, and formulate her advice accordingly. This also means that there are no objective criteria to assess which option should be preferred.

And here we have a problem, because a chatbot has no preference, but is simply programmed. We could hypothetically program preferences in the bot in case of conflicting obligations. We could for instance say that if rule 1 conflicts with rule 2, rule 1 prevails. This would mean, applied in the example above, that the bot should advise to continue the breastfeeding. However, this would also mean that the chatbot would blindly give preference to one moral theory regardless the moral setting. What if the worried mother says that she is considering suicide because she is about to collapse as a result of the constant stress? Would the chatbot then still have to prefer the deontological obligation over the consequentialist one? And if we would hypothetically add another principle (for instance, rule 3: advice may never lead to risking the death of a customer), can we program the chatbot in such a way that it knows exactly when to give priority to rule 1 (right to health) and rule 3 (right to life)? It will be very hard to program a bot in such a way that it knows exactly what to do regardless of contextual input. However, there are some approaches in meta ethics that oppose the idea of incomparability. They can be used to integrate contextual input in solving the problem of moral uncertainty. This way, moral programming matches human needs in machine ethics more adequately. It also enables engineers to use societal/user input in the development of their moral programming, and –if desired-the review of their moral programming against the initial expectations when applied in practise.

4.3.2. *A comparative approach*

Various approaches are proposed to deal with moral uncertainty, such as the comparative approach and the expected moral value approach. In a comparative approach [29], the choice between approaches in normative ethics depends of the aim of the moral actor (see also Fig. 5). If it is the goal of the moral actor to pursuit value X, then the moral action should be the one that most contributes to that goal: *"So the appropriate action corresponding to being better in value which corresponds to be being* supported *by most reason is choosing that option. And the appropriate action corresponding to being worse in value, which corresponds to being* supported *by less reason, is not choosing that option"* [29]; p.116). Chang uses a so-called 'Archimedean point' (some comparatists would use the Latin term 'Tertium Comparationis' instead), which is used as a fixed point with which we can compare different approaches in normative ethics [59,61]. So, in the case of Sylvia, the chatbot working for the infant formula company, this would mean that it depends on which value the company finds the most important, and if it will act accordingly. If the most important value is considered to be children's health, we would have to compare which approach in normative ethics would lead to the most complete fulfilment of this value. As mentioned above, a utilitarianism approach would most likely not result in advising to continue breastfeeding; it would instead advise to look for an

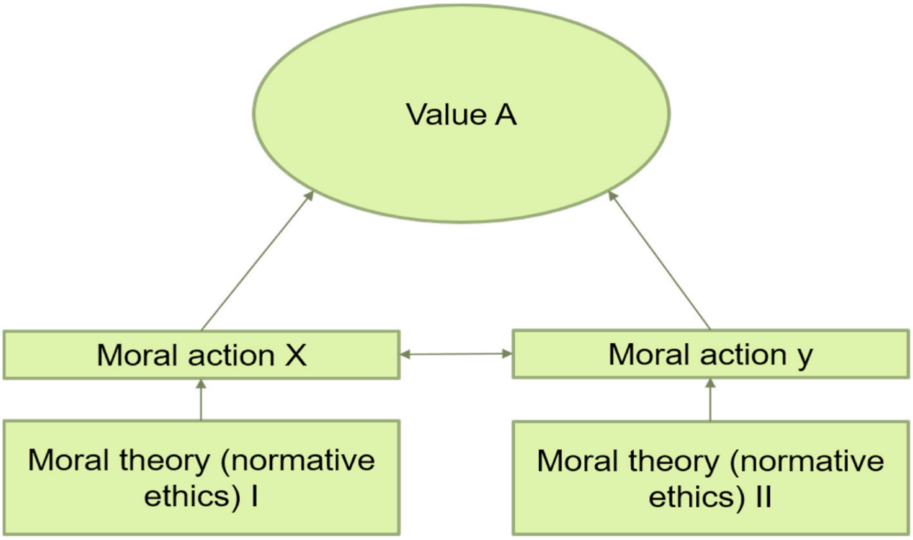
Value A
Moral action X
Moral action y
Moral theory (normative ethics) I
Moral theory (normative ethics) II

Fig. 5. Comparativism in meta ethics.

alternative (such as infant formula). The utilitarian approach would not seriously damage the health of the child, but would not lead to as much fulfilment of the value 'children's health' as a deontological –rights-based- approach would, where breastfeeding is considered a fundamental human right. If the most important value would be family wellbeing, there is certainly reason to muster a utilitarian approach, which would mean an advice to stop breastfeeding and use infant formula instead. In the context of moral programming, this would mean that the programming should include a priority list of values that are important to achieve, and a workable and measurable definition of what these values mean, so that actions can be accounted for using percentages of expected realization of the chosen value. So, if we want to realize value A, and moral theory I leads to action X, which leads to 90% realization of value A, where moral theory II leads to action y, which leads to 60% realization of value A, the most reasonable thing to do is to choose moral action X.

However, an exclusive focus on values as a goal can become problematic since it could easily lead to so-called 'perverse instantiation', where the intention of the programmer or context of the applied action is not taken into due consideration [90] and leads to counterproductive applications. For instance, if Sylvia is programmed to protect the health of the child as a primary value, and notices that the mother who calls is a poor educator with an unhealthy lifestyle, the chatbot might conclude that the child's health may be best protected if it would be raised by foster parents. Such an advice is probably not in line with the expectation of the programmer, nor is it perceived to be appropriate in the context of a B to C relation in which a chatbot is used. To solve this problem, some authors propose to use the Agent-Deed-Consequence (ADC) Model [32,91]. The comparison of something valuable compared to something else that is valuable can be narrowed down to positive and negative evaluations of the moral agent, the moral deed in itself and the consequences of the moral deed, more or less representing elements of virtue ethics, deontology and consequentialism. This could potentially lead to more balanced moral programming. The good deed (protecting the child's health) can be counterbalanced by the consequences (recommending a separation of the child and the mother) and the intention of the programmer (creating a reliable chatbot to give advice in nutrition matters). In other words: the ADC model allows a more nuanced comparison, in which at least three different approaches in normative ethics can be simultaneously used in a comparative setting to define what is most valuable. In this line of reasoning, Aliman and Kester [92,93] propose an ethical framework of augmented utilitarianism, through which society can define ethical goal functions for Artificial Intelligence. The concept of 'the greatest happiness for the greatest number' is now replaced by these goal functions, allowing various ethical perspectives in the equation. In formulating these goal functions, they move away from the debate on what a machine should do, and instead focus on what we want it to do. This can change over time and may differ per cultural setting or even per user. Moral programming should result in the most optimal application of these predefined ethical goal functions, and the output should be permanently reviewed against these goal functions.

In both the ADC model and the framework of augmented utilitarianism, we step away from classic approaches in normative ethics and move towards a more hybrid approach through which various elements of normative ethics can be used in determining and comparing the value of moral programming in the context of utility functions in smart technology. During this process, human intention and perception is constantly taken into consideration, creating more flexible programming for a more accurate translation of human perspectives on ethics into machine ethics.

4.3.3. *The expected moral value approach*

Another workable solution to moral uncertainty is offered by Ted Lockard [94] by introducing the Expected Moral Value Approach. The idea is that a moral actor should take into account not only her own personal belief regarding the credibility of a theory on normative ethics, but also assess to what extent the moral action that is proposed fulfils the values that are supposed to be fulfilled by that moral theory. Let's assume that in a scale of 0,0 to 1.0., the moral authority that decides on how to morally program the technology awards 0,6 credibility to utilitarianism, and 0,4 credibility to a right-based approach to deontology. Let's furthermore assess that in a scale of 0–10, the solution to advise on the continuation of breastfeeding (action A) leads to some happiness for the child in the long term (because of the healthier milk), but also leads to less happiness for the family in general, and rank this with a 2. Let's furthermore assume that the advice to use infant formula (action B) leads to most happiness for most people, but not fully, since in the end, natural breastmilk is healthier for the child: 8.

This leads to the following equations:

Subjective credibility of theory I * Value of action A according to theory I = the desirability of the action

In casu this would be: 0,6 * 2 = 1,2

Subjective credibility of theory I * Value of action B according to theory I = the desirability of the action

In casu this would be: 0,6 * 8 = 4,8

When we do the same for moral theory II (right-based approach to deontology), we could say that advising on giving breast milk is slightly better in order to realize the value health compared to advising on infant formula. However, the infant formula in itself is not unhealthy to the child, and a widely accepted alternative to breastmilk. Let's say that the advice to continue the breastfeeding is the healthiest alternative possible (10) and offering infant formula is not far from this (8). This leads to the following equations:

Subjective credibility of theory II * Value of action A according to theory II = the desirability of the action

In casu this would be: 0,4 * 9 = 3,6

Subjective credibility of theory II * Value of action B according to theory II = the desirability of the action

In casu this would be: 0,4 * 8 = 3,2

Based on this, we can conclude that the chatbot should give the advice to offer infant formula, since action B under theory I has the highest desirability score.

4.3.4. *Flaws and future discussion*

Some scholars reject these approaches on how to deal with the incomparability of normative ethics in itself. Harman [95] for instance, argues that there are serious flaws in these meta-ethical approaches, because they offer solutions that are subjective. After all, there is a certain value judgment in assuming that a moral actor should choose the solution of the highest expected moral value, and that false assumptions of what is right or wrong can be embedded in the consideration when they are genuinely perceived to be true. Sepielli [89] specifically criticizes the Expected Moral Value approach, but does not offer an alternative. He argues that assuming equality between moral theories is a random choice for various reasons. Decisions taken in theory I are not necessarily as important as decisions taken in theory II. Besides that, it is not easy to compare a Kantian principle with a human right, or a number of utils (utilitarianism): the assumed equality is therefore very random, and the value of a moral solution sometimes even depends on the amount of possible moral options per theory. Most of the criticism is aimed at the fact that a meta-ethical view is expected to offer an objective method of comparing and evaluating normative ethics. Both the comparative approach and the Expected Moral Value Approach do not result in such objective methods. However, in the application in the moral programming of technology, these flaws become less relevant. After all, in the ethical programming of a machine, we are not necessarily looking for a way to asses normative ethics in an objective way, but we are looking for a method to combine various approaches in normative ethics in moral programming, without the risk that the programming leads to conflicting results. This means that we need a way to make meta ethics measurable so we can account for moral choices in the moral programming of technology. Both the comparative approach and the Expected Moral Value approach might just do that.

5. Solutions: towards a roadmap

As stated above, what is lacking so far is a structuralized discourse on moral programming in a smart society. In this contribution, we propose a roadmap that defines how ethical decisions in moral programming can be done. The synopsis of the above is visualized in Fig. 6, reformulated in four questions.

1. *When* do we need to use this ethical roadmap?
2. *Who* has the moral authority to solve a moral problem through moral programming?
3. *What* can be decided when we need to solve a moral problem through moral programming?
4. *How* can the moral judgment be programmed in the smart technology, especially when more than one ethical solution seems to be able to solve the moral problem?

Step 1 (*when*): The roadmap is useful in contexts in which smart technology is used in programmed ethical decision making to solve a moral problem. A moral problem can be solved by more than one (mutually exclusive) value judgment. When programming involves dealing with a moral problem, we need a structuralized ethical decision-making model that suits the needs of an engineer who is responsible for the functioning of the smart technology.

Step 2 (*who*): To this end, the engineer should first be aware of the moral authority in a given moral problem. We propose that in a low stake moral problem, the engineer is the most suitable authority to decide on the nature of the programming, due to the limited effects for the involved moral stakeholders. However, in a high-stake moral problem, we propose that the individual user or society (the legislator) is the most suitable moral authority to decide on moral programming, depending on the concentration of the effects of the programming. When the consequences are individual, concentrating on the user, we assume that the consumer that interacts with the technology is the moral authority. When the consequences are more societal, the moral authority should be a legislator, formalizing the solution of the ethical dilemma through a legislative process. To be able to qualify a moral problem as a low-stake or high-stake problem, as well as to be able to recognize the concentration of the effect, we propose to use the moral intensity dashboard (Fig. 2) that is based on Jones' [18] theory on moral intensity, analysing the issue-related aspects of a moral dilemma. This dashboard can help the engineer in assessing which moral stake corresponds with which moral issue, and determine who has the moral authority to decide on the nature of the moral programming.

Step 3 (*what*): A value judgment can be supported by various approaches in normative ethics. We have argued that universalist and absolutist theories in normative ethics are typically suitable for moral programming. Smart technology may act morally but cannot be moral in itself. Therefore, we need to program technology that uses normative ethics that is clearly defined outside the scope of human characteristics which involves the capacity to have subjective awareness. This rules out the more relativist normative ethical approaches that would require a P-conscious mind: something a machine in the foreseeable future will not (yet) have. This leaves consequentialist and deontological theories as suitable starting points for moral programming in itself. This means that the ethics we expect machines to apply can best be formulated as goals or principles.

Step 4 (*how*): Deontic logic seems to be the most straightforward way to morally program technology, using the terminology of Von Wright into practice. This means that moral programming should be built around four action types: obligatory (o), permitted (p), forbidden (f) and indifferent (i) actions. One of the characteristics of a moral problem is that different approaches in normative ethics lead to opposing conclusions. Therefore, meta-ethical programming is required to enable smart technology that is confronted with a moral problem to perform an inter-theoretic value comparison and select the most suitable ethical solution. There are various approaches that could offer a valid starting point. In a comparative approach, the programming is focused on prioritized measurable values, comparing effects of various ethical solutions, and selecting the most effective one towards the prioritized value. Both the Agent-Deed-Consequence model and the framework of augmented utilitarianism offer methods that enable programmers to combine various approaches in normative ethics to avoid an exclusive focus on values as a goal in itself. In the Expected Moral Value Approach, we do not prioritize values, instead we prioritize approaches in normative ethics and include this in the equation. In the equation itself, we calculate the value of an action according to one theory and compare that to the value of another action according to the same theory. We can then apply a similar process using another ethical theory and calculate the most suitable solution to the moral problem. In both approaches, we need continuous and organized input from the moral authority (the engineer, the individual user or society) to define the preferences (e.g. ethical goal functions or the prioritizing of ethical approaches). This can lead to a flexible and ever-improving input for moral programming that does justice to the relevant human perceptions of what we want smart

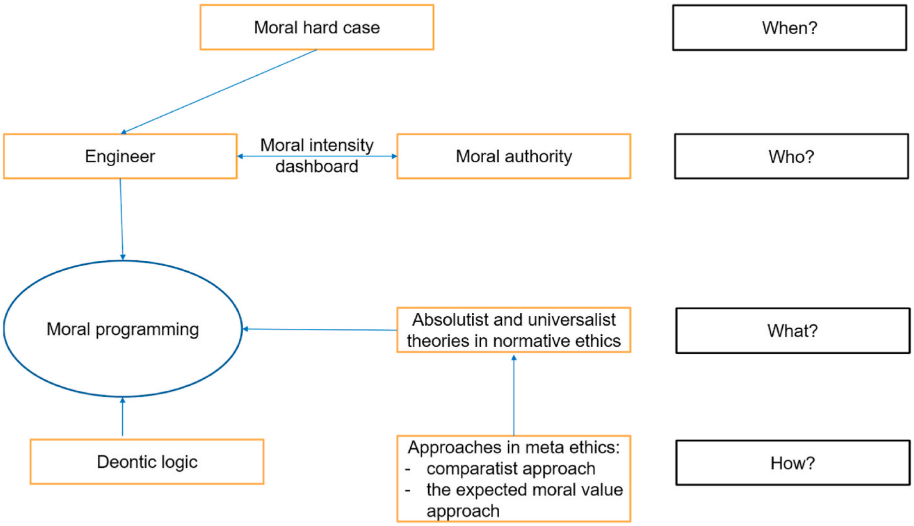
Moral hard case
When?
Engineer
Moral intensity dashboard
Moral authority
Who?
Moral programming
Absolutist and universalist theories in normative ethics
What?
Deontic logic
Approaches in meta ethics:
- comparatist approach
- the expected moral value approach
How?

Fig. 6. Roadmap for the moral programming of smart technology.

technology to do as a moral agent.

6. Implications and conclusion

The role and impact of smart technology in our society is rapidly increasing. One of the main implications of our conceptual exposition is that smart technology can only follow an absolutist/universalist process, while human ethical decision making appears to be more of a relativist nature. Consequentially, machine ethics differs from human ethics. This implies that machine ethics cannot be presented as an alternative to human ethics or misrepresented as being the same. In the interaction between humans and smart technology this has one obvious implication: when humans are not reasonably aware of the fact that they interact with a machine that is ethically programmed, they should be informed. For instance, humans could be confused when companies would run experiments to make chatbots sound and act as humanly as possible, or even mimic human-inspired characteristics and emotions. When such a human-sounding chatbot makes a value judgment, the consumer should be aware of the ethical nature of such a judgement: machine, and not human.

In this contribution we offer a roadmap to support the ethical decision-making process in moral programming. As always, when proposing some sort of formalized structure in taking ethical decisions (EDP), *'no single model is universally accepted'* [18], it will come as no surprise that the roadmap we propose is a starting point; a contribution to a current academic discourse, that asks for further research, most probably of a more empirical nature. The next step should be twofold. First, we should validate whether engineers can work with a moral intensity dashboard and validate the functionality dimensions in moral intensity in various professional settings. Second, we need to generate societal and user input in ethical high-stake settings. In other words: we need to know what we should put in the program when the engineer is not the moral authority. This requires empirical research into moral data: we need to know what people would prefer in the ethical programming in a given moral problem. This 'moral data' can be generated in various test labs and should result in practical input that engineers can work with. On top of that, we need to establish a feedback loop to establish whether the predicted output of the moral programming indeed matches the human expectations [96].

Author statement

BW is the sole author of this conseptual paper, and fully acknowledged as author of the entire paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.techsoc.2020.101466.

References

[1] C.E. Jimenez, A. Solanas, F. Falcone, E-government interoperability: linking open and smart government, IEEE Comput. 47 (10) (2014) 22–24, https://doi.org/10.1109/MC.2014.281.

[2] T. Mazali, From industry 4.0 to society 4.0, there and back, AI Soc. 33 (3) (2018) 405–411, https://doi.org/10.1007/s00146-017-0792-6.

[3] M.J. Sousa, Á. Rocha, Skills for disruptive digital business, J. Bus. Res. 94 (2019) 257–263, https://doi.org/10.1016/j.jbusres.2017.12.051.

[4] P. Boddington, Towards a Code of Ethics for Artificial Intelligence, Springer, Cham, 2017.

[5] M. Martínez-Córcoles, M. Teichmann, M. Murdvee, Assessing technophobia and technophilia: development and validation of a questionnaire, Technol. Soc. 51 (2017) 183–188, https://doi.org/10.1016/j.techsoc.2017.09.007.

[6] E. Awad, S. Dsouza, R. Kim, et al., The moral machine experiment, Nature 563 (2018) 59–64, https://doi.org/10.1038/s41586-018-0637-6.

[7] P. Lin, R. Jenkins, K. Abney, Robot Ethics 2.0, Oxford University Press, Oxford, 2017 (edits).

[8] P.J.M. Elands, A.G. Huizing, L.J.H.M. Kester, M.M.M. Peeters, S. Oggero, Governing ethical and effective behaviour of intelligent systems, A novel framework for meaningful human control in a military context, Mil. Spect. 188 (6) (2019) 302–313.

[9] S. Nyholm, J. Smids, The ethics of accident-algorithms for self-driving cars: an applied Trolley problem? Ethical Theory & Moral Pract. 19 (2016) 1275–1289, https://doi.org/10.1007/s10677-016-9745-2.

[10] R. Sparrow, Robots, rape and representation, Int. J. Soc. Robot. 2017 (9) (2017) 465–477, https://doi.org/10.1007/s12369-017-0413-z.

[11] L. Zardiashvili, E. Fosch-Villaronga, ''Oh, dignity too?'' said the robot: human dignity as the basis for the governance of robotics, Minds Mach. (2020) 1–23, https://doi.org/10.1007/s11023-019-09514-6.

[12] S. Glende, I. Conrad, L. Krezdorn, S. Klemcke, C. Krätzel, Increasing the acceptance of assistive robots for older people through marketing strategies based on stakeholder needs, Int. J. Soc. Robot. 8 (2016) 355–369, https://doi.org/10.1007/s12369-015-0328-5.

[13] C.L. Miltgen, J. Henseler, C. Gelhard, A. Popovič, Introducing new products that affect consumer privacy: a mediation model, J. Bus. Res. 69 (10) (2016) 4659–4666, https://doi.org/10.1016/j.jbusres.2016.04.015.

[14] B. Roessler, Should personal data be a tradeable good? On the moral limits of markets in privacy, in: B. Roessler, D. Mokrosinska (Eds.), Social Dimensions of Privacy: Interdisciplinary Perspectives, Cambridge University Press, Cambridge, 2015, pp. 141–161.

[15] A.S. Thorpe, S. Roper, The ethics of gamification in a marketing context, J. Bus. Ethics 155 (2) (2017) 597–609, https://doi.org/10.1007/s10551-017-3501-y.

[16] S.A. Wright, A.E. Schultz, The rising tide of artificial intelligence and business automation: developing an ethical framework, Bus. Horiz. 61 (6) (2018) 823–832, https://doi.org/10.1016/j.bushor.2018.07.001.

[17] M.S. Schwartz, Ethical decision-making theory: an integrated approach, J. Bus. Ethics 139 (4) (2016) 755–776, https://doi.org/10.1007/s10551-015-2886-8.

[18] T.M. Jones, Ethical decision making by individuals in organizations: an issue-contingent model, Acad. Manag. Rev. 16 (2) (1991) 366–395.

[19] J.R. Rest, Moral Development: Advances in Research and Theory, Praeger, New York, 1986.

[20] L.K. Trevino, Ethical decision making in organizations: a person-situation interactionist model, Acad. Manag. Rev. 11 (3) (1986) 601–617, https://doi.org/10.2307/258313.

[21] J.L. Craft, A review of the empirical ethical decision-making literature: 2004-2011, J. Bus. Ethics 117 (2) (2013) 221–259, https://doi.org/10.1007/s10551-012-1518-9.

[22] M.J. O'Fallon, K.D. Butterfield, A review of the empirical ethical decision-making literature: 1996–2003, J. Bus. Ethics 59 (4) (2005) 375–413, https://doi.org/10.1007/s10551-005-2929-7.

[23] T.W. Loe, L. Ferrell, P. Mansfield, A review of emperical studies assessing ethical decision making in business, J. Bus. Ethics 25 (3) (2000) 185–204, https://doi.org/10.1023/A:1006083612239.

[24] R.C. Ford, W.D. Richardson, Ethical decision making: a review of the empirical literature, J. Bus. Ethics 13 (3) (1994) 205–221.

[25] J. Doorn, M. Mende, S. Noble, J. Hulland, A. Ostrom, D. Grewal, J. Petersen, Domo arigato Mr. Roboto: emergence of automated social presence in organizational frontlines and customers' service experiences, J. Serv. Res. 20 (1) (2017) 43–58, https://doi.org/10.1177/1094670516679272.

[26] D. Statman, The debate over the so-called reality of moral dilemmas, Phil. Pap. 19 (3) (1990) 191–211, https://doi.org/10.1080/05568649009506337.

[27] W. Dubbink, A typology of ethical problems, Ethical Perspect. 25 (4) (2018) 683–714.

[28] W.D. Ross, The Foundations of Ethics, Oxford University Press, Oxford, 1939.

[29] R. Chang, Are hard choices cases of incomparability? Phil. Issues 22 (2012) 106–126, https://doi.org/10.1111/j.1533-6077.2012.00239.x.

[30] M. Messerli, K. Reuter, Hard cases of comparison, Phil. Stud. 174 (2017) 2227–2250, https://doi.org/10.1007/s11098-016-0796-y.

[31] J. Millar, Ethics settings for autonomous vehicles, in: P. Lin, R. Jenkins, K. Abney (Eds.), Robot Ethics 2.0, Oxford University Press, Oxford, 2017, pp. 20–34.

[32] V. Dubljević, Toward Implementing the ADC Model of Moral Judgment in Autonomous Vehicles, Science and Engineering Ethics, 2020, https://doi.org/10.1007/s11948-020-00242-0.

[33] J. Millar, Technology as moral proxy: autonomy and paternalism by design, in: 2014 IEEE International Symposium on Ethics in Science, Technology and Engineering, Chicago, U.S.A, 2014, pp. 1–7, https://doi.org/10.1109/ETHICS.2014.6893388.

[34] J. Auernhammer, Human-centered AI: the Role of Human-Centered Design Research in the Development of AI. DRS2020, August 2020, 2020, https://doi.org/10.21606/drs.2020.282. Brisbane, Australia.

[35] S. Leavy, Gender bias in artificial intelligence: the need for diversity and gender theory in machine learning, in: Proceedings of the 1st International Workshop on Gender Equality in Software Engineering (GE '18). ACM, New York, 14-16, 2018.

[36] M. Garcia, Racist in the machine: the disturbing implications of algorithmic bias, World Pol. J. 33 (4) (2016) 111–117.

[37] P. Verbeek, Materializing morality, design ethics and technological mediation, Sci. Technol. Hum. Val. 31 (3) (2006) 361–379, https://doi.org/10.1177/0162243905285847.

[38] A. Van Wynsberghe, Health Care Robots, Ethics, Design and Implementation, Routeledge, London-New York, 2016.

[39] J. Borenstein, R.C. Arkin, Nudging for good: robots and the ethical appropriateness of nurturing empathy and charitable behaviour, AI Soc. 32 (2017) 499–507, https://doi.org/10.1007/s00146-016-0684-1.

[40] S.T. Hannah, B.J. Avolio, D.R. May, Moral maturation and moral conation: a capacity approach to explaining moral thought and action, Acad. Manag. Rev. 36 (4) (2011) 663–685, https://doi.org/10.5465/amr.2010.0128.

[41] S.J. Reynolds, Moral attentiveness: who pays attention to the moral aspects of life? J. Appl. Psychol. 93 (5) (2008) 1027–1041, https://doi.org/10.1037/0021-9010.93.5.1027.

[42] P.H. Werhane, Moral imagination and the search for ethical decision-making in management, Bus. Ethics Q. 8 (1998) 75–98.

[43] J. Haidt, The emotional dog and its rational tail: a social intuitionist approach to moral judgement, Psychol. Rev. 4 (2001) 814–834, https://doi.org/10.1037/0033-295X.108.4.814.

[44] H.S. James Jr., Reinforcing ethical decision making through organizational structure, J. Bus. Ethics 28 (2000) 43–58, https://doi.org/10.1023/A:1006261412704.

[45] W.T. Ross, D.C. Robertson, A typology of situational factors: impact on salesperson decision-making about ethical issues, J. Bus. Ethics 46 (2003) 213–234, https://doi.org/10.1023/A:1025563624696.

[46] W. Crane, D. Matten, Business Ethics, 4t edition, Oxford University Press, Oxford, 2016.

[47] J.M. McMahon, R.J. Harvey, The effect of moral intensity on ethical judgment, J. Bus. Ethics 72 (2007) 335–357, https://doi.org/10.1007/s10551-006-9174-6.

[48] K.D. Butterfield, L.K. Trevino, G.R. Weaver, Moral awareness in business organizations: influences of issue-related and social context factors, Hum. Relat. 53 (7) (2000) 981–1018, https://doi.org/10.1177/0018726700537004.

[49] G. Kent, Child feeding and human rights, Int. Breastfeed. J. 1 (27) (2006) 1–12, https://doi.org/10.1186/1746-4358-1-27.

[50] International Code of Marketing of Breast-milk Substitutes, World Health Organization, Geneva, 1981.

[51] S. Ip, M. Chung, G. Raman, et al., Breastfeeding and maternal and infant health outcomes in developed countries, Evid. Rep. Technol. Assess. 153 (2007) 1–186.

[52] R.K. Mitchell, B.R. Agle, D.J. Wood, Towards a theory of stakeholder identification and salience: defining the principle of who and what really counts, Acad. Manag. Rev. 22 (4) (1997) 853–886.

[53] M.A. Mayo, L.J. Marks, An empirical investigation of a general theory of marketing ethics, J. Acad. Market. Sci. 18 (1990) 163–171, https://doi.org/10.1007/BF02726432.

[54] S.D. Hunt, S.J. Vitell, A general theory of marketing ethics, J. Macromarketing 5 (1) (1986) 5–16, https://doi.org/10.1177/027614678600600103.

[55] S.D. Hunt, S.J. Vitell, The general theory of marketing ethics, a revision and three questions, J. Macromarketing 26 (2) (2006) 143–153, https://doi.org/10.1177/0276146706290923.

[56] D.R. May, K.P. Pauli, The role of moral intensity in ethical decision making, Bus. Soc. 41 (1) (2002) 84–117, https://doi.org/10.1177/0007650302041001006.

[57] Y. Yu, Comparative analysis of Jones' and Kelly's ethical decision-making models, J. Bus. Ethics 130 (3) (2015) 573–583, https://doi.org/10.1007/s10551-014-2245-1.

[58] S.J. Harrington, A test of a person – issue contingent model of ethical decision making in organizations, J. Bus. Ethics 16 (4) (1997) 363–375, https://doi.org/10.1023/A:1017900615637.

[59] B. Wernaart, Ethics and Business, a Global Introduction, Groningen/Houten: Noordhoff Uitgevers, 2018.

[60] I. Kant, Groundwork for the Metaphysic of Morals, Yale University Press, London, 2002 original publication: Riga, 1785.

[61] B. Wernaart, The Enforceability of the Human Right to Adequate Food, a Comparative Study, Wageningen Academic Publishers, Wageningen, 2013.

[62] J. Rawls, A Theory of Justice, Harvard University Press, Cambridge, 1971.

[63] J. Rawls, Justice as Fairness, a Restatement, Harvard University Press, Cambridge, 2001.

[64] T. Tännsjö, Moral relativism, Phil. Stud. 135 (2) (2007) 123–143.

[65] A. MacIntyre, After Virtue, a Study in Moral Theory, Notre Dame: University of Notre Dame Press, 1981.

[66] C.W. Gowans, Virtue ethics and moral relativism, in: Hales (Ed.), A Companion to Relativism, Blackwell Publishing Ltd, Oxford, 2011, pp. 391–410.

[67] Z. Bauman, Postmodernist Ethics, Blackwell Publishing, Oxford, 1993.

[68] R. Rorty, Philosophy and the Mirror of Nature, Princeton University Press, 1979.

[69] C. Gilligan, In a Different Voice, Harvard University Press, Cambridge, 1982.

[70] J.L. Borgerson, On the harmony or feminist ethics and business ethics, Bus. Soc. Rev. 112 (4) (2007) 477–509, https://doi.org/10.1111/j.1467-8594.2007.00306.x.

[71] J. Habermas, Moral Consciousness and Communicative Action, MIT Press, Cambridge, 1992.

[72] G. Demuijnck, Universal values and virtues in management versus cross-cultural moral relativism: an educational strategy to clear the ground for business ethics, J. Bus. Ethics 128 (2015) 817–835, https://doi.org/10.1007/s10551-014-2065-3.

[73] M. Nussbaum, Non-relative virtues: an Aristotelian approach, Midwest Stud. Philos. 13 (1988) 32–50, https://doi.org/10.1093/0198287976.003.0019.

[74] A.G. Scherer, M. Patzer, Beyond Universalism and Relativism: Habermas's Contribution to Discourse Ethics and its Implications for Intercultural Ethics and Organization Theory, 2011, April 28, https://doi.org/10.1108/S0733-558X(2011)0000032008. IOU Working Paper No. 119.

[75] M. Johnson, Moral Imagination, Implications of Cognitive Science for Ethics, The University of Chicago Press, Chicago, 1993.

[76] J. Thomson, The Trolley problem, Yale Law J. 94 (6) (1985) 1395–1415.

[77] P. Lin, Why ethics matters for autonomous cars, in: M. Maurer, et al. (Eds.), Autonomes Fahren, Springer, Berlin, 2015, pp. 69–85, https://doi.org/10.1007/978-3-662-48847-8_4.

[78] K. Arkoudas, S. Bringsjord, S. Bello, Toward ethical robots via mechanized deontic logic, in: Machine Ethics, Papers from the Fall AAAI Symposium; FS-05-06, American Association for Machine Intelligence, 2005, pp. 17–23.

[79] T.M. Powers, Prospects for a Kantian Machine, Intelligent Systems, IEEE, 2006, pp. 46–51, https://doi.org/10.1109/MIS.2006.77. August 2006.

[80] G.H. von Wright, Deontic logic, Mind LX (237) (1951) 1–15.

[81] S. Bringsjord, A 21st-century ethical hierarchy for robots and persons: EH, in: Ferreira, et al. (Eds.), A World with Robots, International Conference on Robot Ethics: ICRE 2015, Springer, Cham, 2017, pp. 47–61.

[82] T. Hedberg, Epistemic supererogation and its implications, Synthese 191 (15) (2014) 3621–3637, https://doi.org/10.1007/s11229-014-0483-5.

[83] J.O. Urmson, Saints and heroes, in: A.I. Melden (Ed.), Essays in Moral Philosophy, university of Washington Press, Seattle, 1958, pp. 198–216.

[84] N. Block, On a confusion about a function of consciousness, Behav. Brain Sci. 18 (1995) 227–287.

[85] S. Bringsjord, Offer: one billion dollars for a conscious robot. If you're honest, you must decline, J. Conscious. Stud. 14 (7) (2007) 28–43.

[86] A.M. Johnson, S. Axinn, Acting vs. being moral: the limits of technological moral actors, in: IEEE International Symposium on Ethics in Science, Technology and Engineering, 22-24 May 2014, 2014, https://doi.org/10.1109/ETHICS.2014.6893396.

[87] G.J.C. Lokhorst, Computational meta-ethics, towards the meta-ethical robot, Minds Mach. 21 (2011) 261–274, https://doi.org/10.1007/s11023-011-9229-z.

[88] J. Raz, Value incommensurability: some preliminaries, Proc. Aristot. Soc. 86 (1) (1986) 117–134.

[89] A. Sepielli, Moral uncertainty and the principle of equity among moral theories, Philos. Phenomenol. Res. 86 (3) (2008) 580–589, https://doi.org/10.1111/j.1933-1592.2011.00554.x.

[90] R.V. Yampolskiy, Artificial intelligence safety engineering: why machine ethics is a wrong approach, in: V. Müller (Ed.), Philosophy and Theory of Artificial Intelligence. Studies in Applied Philosophy, Epistemology and Rational Ethics, vol. 5, Springer, Berlin, 2013, pp. 389–396, https://doi.org/10.1007/978-3-642-31674-6_29.

[91] V. Dubljević, E. Racine, The ADC of moral judgment: opening the black box of moral intuitions with heuristics about agents, deeds and consequences, AJOB Neurosci. 5 (4) (2014) 3–20, https://doi.org/10.1080/21507740.2014.939381.

[92] N. Aliman, L. Kester, Augmented Utilitarianism for AGI Safety, 2019 arXiv:1904.01540.

[93] N. Aliman, L. Kester, Requisite Variety in Ethical Utility Functions for AI Value Alignment, 2019 arXiv:1907.00430.

[94] T. Lockard, Moral Uncertainty and its Consequences, Oxford University Press, Oxford, 2000.

[95] E. Harman, The irrelevance of moral uncertainty, in: R. Shafer-Landau (Ed.), Oxford Studies in Metaethics, ume 10, Oxford University Press, Oxford, 2015.

[96] N. Aliman, L. Kester, Transformative AI governance and AI-empowered ethical enhancement through preemptive simulations, Delphi – Interdiscipl. Rev. Emerg. Technol. 2 (1) (2019) 23–29, https://doi.org/10.21552/delphi/2019/1/6.