religions

MDPI

Article

The Limits of Machine Ethics

Sara Lumbreras

Institute for Research in Technology, Universidad Pontificia Comillas, Madrid 28001, Spain; slumbreras@comillas.edu

Academic Editor: Noreen Herzfeld
Received: 17 March 2017; Accepted: 17 May 2017; Published: 19 May 2017

Abstract: Machine Ethics has established itself as a new discipline that studies how to endow autonomous devices with ethical behavior. This paper provides a general framework for classifying the different approaches that are currently being explored in the field of machine ethics and introduces considerations that are missing from the current debate. In particular, law-based codes implemented as external filters for action—which we have named *filtered decision making*—are proposed as the basis for future developments. The emergence of values as guides for action is discussed, and personal language –together with subjectivity- are indicated as necessary conditions for this development. Last, utilitarian approaches are studied and the importance of objective expression as a requisite for their implementation is stressed. Only values expressed by the programmer in a public language—that is, separate of subjective considerations—can be evolved in a learning machine, therefore establishing the limits of present-day machine ethics.

Keywords: ethics of machines; theory of mind; values; learning automata.

1. Machine Ethics: The Rise of a New Field

Robots are growing in importance; they are assuming more roles in the industry and are expected to be an expanding factor in the future economy (International Federation of Robotics IFR 2016). In addition, the interactions between robots and humans are increasingly intense, even appearing in some sensitive fields such as military activities or medical caregiving. This has led some authors to describe the next century as the *age of robots* (Veruggio et al. 2016, pp. 2135–60). These human–machine interactions are already reshaping the way we relate to each other and process information, in some favorable as well as detrimental ways (Head 2014).

This has motivated the emergence of *roboethics*, which deals with the issues derived from the diverse applications of robots understood as particular devices[1] and of *machine ethics*, which studies how to endow self-regulated machines with ethical behavior. It should be noted that there is no consensus yet about the scope of these two fields, and the previously presented definitions have been chosen because of their usefulness and are being accepted by a seemingly increasing group of scholars (Veruggio et al. 2016, pp. 2135–60). There have been myriad works dealing with roboethics and exploring the implications of the use of robots in particular activities (Luxton 2014, pp. 1–10), such as their implications for employment, but the contributions to the ethics of machines are still taking shape. The first efforts focused on the ethics of automatic machines insofar as their actions impacted human beings. With this aim, some authors have classified the interactions into mere observation, interference, interaction, or advice (Van de Voort et al. 2015, pp. 41–56). However, there have been

[1] An autonomous device is understood as a machine that receives some sensory *inputs* from the environment and, using some specific rules, reaches an *action* that can have an impact on the environment. The term *autonomous device* (or, *device*, for short) is used in this text interchangeably with *robot*, *machine*, *autonomous machine*, or *automaton*. A *learning* automaton would be the automaton where these rules are not static but rather change to adapt to achieve a given defined goal.

some voices claiming for the dignity of the machines in themselves rather than just objects having an impact on humans, such as the *ontocentric* ethics proposed by Floridi (Floridi 2005, p. 3).

This paper attempts to provide a comprehensive framework for the different proposals that have appeared in the context of machine ethics and discusses their relative worth. Its main contributions are as follows:

> Providing a comprehensive framework for analyzing the different proposals within machine ethics. This framework is based in the existing double dichotomy negative vs. positive and top-down vs. bottom-up. The paper studies their relative strengths and disadvantages.
>
> Proposing a mechanism to ensure compliance in evolving, interconnected devices, which has been named *filtered decision making*.
>
> Identifying subjectivity (understood as *being a subject, an individual which possesses conscious experiences, such as perspectives, feelings, beliefs and desires* (Honderich 2005)) as the final frontier of machine ethics. Approaches that require it should not be considered implementable with present-day technology, as there is no basis to suppose that subjectivity and consciousness will emerge from it.
>
> Analyzing the conditions for the implementation of utilitarian approaches, which establish a further limit for machine ethics: the objective function of the automaton should be expressed in a public language in order to be implementable.

The rest of this paper is organized as follows. First, the main existing proposals for machine ethics are classified into a double categorization of *negative* vs. *positive* or *bottom-up* vs. *top-down*. This is briefly presented in Section 2. Then, negative approaches are discussed in depth in Section 3, with Section 4 focusing on filtered decision making, one of the main contributions of this paper. Sections 5 and 6 deal with positive approaches and present our conditions for the development of utilitarian approaches. Finally, Section 7 extracts conclusions.

2. A Wide Spectrum of Proposals

The proposals that have been put forward in the context of machine ethics—that is, the diverse procedures that have been proposed for endowing machines with ethical behavior—reflect, quite naturally, the different existing understandings about ethics as a whole.

It is particularly interesting for our purposes to work with the following double dichotomy, which classifies approaches to ethics based on their object:

> *Negative ethics* is concerned with preventing harm to other beings. In general, they can be expressed as moral codes composed of actions that should be avoided (such as killing, stealing, or lying (Howard and Korver 2008; Lichtenberg 2010, pp. 557–78)).
>
> *Positive ethics* focuses on creating the largest good instead of avoiding harm. They are usually consequentialist approaches, where the good associated to a given decision determines if it is the best course of action or not (Handelsman et al. 2009, pp. 105–13; Jackson 1991, pp. 461–82; Slote 1985).

In addition, it is useful to distinguish the way in which these perspectives are developed (Allen et al. 2005, pp. 149–55):

> *Top-down ethics* conceive the moral rules or the definition of ethical good as something objective that is accepted by the agent. Kantian ethics would be an example of this kind of approach (Allen et al. 2005, pp. 149–55; Powers 2006, pp. 46–51; Kant and Abbott 2004).
>
> *Bottom-up ethics* consider that it is the subject who selects the values that guide her behavior, progressively refining them in a learning process that depends on experience.

> An example of this is the plastic definition of values that appears in the work of Campbell (Campbell et al. 2002, pp. 795–823) and in the basis of social cognitive theory (Martin 2004, pp. 135–45) as well as some approaches to morality and identity development in child psychology (Lapsley and Narvaez 2006; Hardy and Carlo 2011, pp. 212–18).

These two dichotomies can be combined with the aim of framing a particular approach to ethics. *Top-down negative ethics* would describe settings where moral rules are externally imposed and determine behaviors that should be avoided. *Top-down positive ethics* present a framework where a desirable output must be maximized but where the definition of what is desirable or not has been externally given. Complementarily, *bottom-up negative ethics* would describe a scheme where the definition of harmful actions to avoid emerges from the experience of the moral subject, while *bottom-up positive ethics* would maximize goodness and discover the meaning of goodness itself.

The following sections make use of these definitions as a way of structuring the approaches that have been proposed in the context of machine ethics[2].

3. Negative Approaches

Negative approaches are undoubtedly appealing as they have been the fundamental approach to ethics until the past century. However, as stated by several authors (Veruggio et al. 2016, pp. 2135–2160), any approach based on rules can result in conflicts (i.e., situations where complying with one rule implies breaking another). This problem was overcome by Asimov by establishing priorities among rules in his rules of robotics (Asimov 1950). For illustration, the first text where the rules appeared is reproduced below:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

It should be noted that establishing priorities does not necessarily solve any conflict generated in rule-based approaches. In some cases, meta-priorities or the random selection of an action (known as *dice throwing*) might be necessary.

Rules can correspond to the expression of a moral duty (such as the categorical imperative in Kant), which implies a considerable difficulty in expressing them in a simple action approval/dismissal framework. In addition, rules can be the reflection of a social contract, of abiding the explicit norms (law and regulations) or implicit rules (customs) of behavior. The second type is more easily amenable to objective expression, which would have considerable advantages when considering implementation into a self-regulating, evolving machine.

It should be noted that opting for a law-based approach does not necessarily entail a modest view of the question of machine ethics. In many fields, unfortunately just abiding the law would result in

---

[2] Dual process theory has been successfully used to explain some of the characteristics of ethical reasoning, and it is worth discussing its merits. The main idea of its proponents is that ethical reasoning happens through two distinct mechanisms: emotion-based, intuitive judgment (1), and rationality-based judgments (2). Greene linked them to deontological ethics and consequentialism, respectively (Greene et al. 2001, pp. 2105–8). The two types of thought according to Kahneman (Kahneman 2011) have been linked to these two mechanisms, with System 1 ("thinking fast") corresponding to emotion-based judgments and System 2 ("thinking slow") describing rationality-based ethical reasoning. Although it might be tempting to identify this dichotomy with top-down vs. bottom-up ethics, they are not exactly equivalent. The main difference is that rationality-based judgment, which as discussed would be related to consequentialism, should be identified with positive ethics in the sense of choosing the action that delivers the largest good. By contrast, bottom-up ethics refers to the emergence of values. Values are implicit in the consequentialist evaluation (which, for instance, in the trolley problem could be expressed as the number of people that remain alive). Consequentialism takes this definition of "goodness of an outcome" as a given rather than provide it as an output as would be the case with the emergence of values in bottom-up ethics. However, the point of combining two different modes of ethical reasoning is extremely powerful and is also one of the features of the framework for machine ethics provided in this paper.

more ethical behavior than the one commonly displayed by human beings. This would be the case, for instance, in war operations, where environmental conditions mean that it is relatively difficult to respect the established rules of behavior.

There have already been some authors, such as Casey (Casey 2017), that claim that law-based approaches are sufficient to tackle the question of machine ethics. This is not the position defended in this paper; however, law-based approaches can be a good starting point for more sophisticated systems and, in any case, a necessary step to ensure safety for human beings in their ever increasing interaction with the automata. The power of the approach proposed would be to combine a negative and a positive approach to ethics of machines as will be described in the rest of the paper.

It should not be forgotten that it is necessary to question our ability to ensure compliance in evolving, learning automata. Safety engineering is, in this context, a growing concern, as robots can easily learn to circumvent the rules if that leads to a more beneficial result (Yampolskiy and Fox 2013, pp. 217–26). A particularly clear example of this is anti-discrimination laws in insurance. While regulations require companies to offer equal treatment to their customers regardless of sex or race, as stated by Ronen, "(insurance companies) are in the business of discriminating" (Avraham et al. 2012). Learning algorithms that have been trained to predict default rates without being given sex or race as an input are able to build proxies of these attributes and use them in their internal calculations, so that their output is more accurate—albeit discriminatory. Knowledge extraction and ex-post testing can be used for this purpose.

4. Within the Limits of What is Possible: Filtered Decision-Making and Transparency

The issue with learning automata is fundamentally a conflict between the top-down nature of law-based ethics with the underlying constructivist mechanics of learning, evolving algorithms. This paper distinguishes two different types of problems related to this conflict:

> It is not obvious to determine whether a learning automaton complies necessarily with some given rules, given that their behavior can be unpredictable even for their programmer.
>
> In addition, learning-based algorithms are not necessarily based on clearly understandable strategies. This might lead to situations where it is unclear whether they comply with top-down ethics or not. For instance, an algorithm might be designed to give quotes for insurance. It is in principle possible for it to evolve towards giving worse quotes to a racial minority, for example, which would violate anti-discrimination laws as explained above. However, if the algorithm is designed in a nontransparent way (for instance, using deep learning[3]), it can be challenging to detect whether this will occur.

I would like to propose *filtered decision making* as a possible solution. This idea consists in designing the learning-based automaton in such a way that it can obtain inputs from the external world and process them to get a decision, but not activate that decision by itself. Instead of determining the output of the process directly, the learning automaton would send the preliminary output to an external, objective module, which would be accessed online. This objective module would assess, in an independent way, whether the action selected by the automaton complies with the established rules. If it does, then the decision is sent back to the action module of the automaton. If it does not, then the automaton goes to its *default action* (determined as the safest by the same committee that designed the top-down rules). In addition, a flag is raised so that the programmer can inspect the issue.

This filtered decision making ensures that the robot will comply with the rules, and eliminates some of the problems that make for tedious implementation in an actual context. General learning machines can be divided into categories depending on their function, for which different sets of rules

---

[3] It should be noted that the particular learning algorithm that is implemented in each case is not relevant, but only the fact that it is not transparent. Any non-transparent machine learning algorithm could have been used in this example.

can be enacted (for instance, trading machines, war automata, and caregiver automata). Moreover, it would be automatic to update these codes when necessary—avoiding the need for updating the programming of the full array of devices in operation. In addition, the filtering module should not be capable of learning, which would ensure it will function correctly in the future.

Filtered decision making draws on the idea of a moral governor that was proposed by Arkin to control lethal behavior in the particular case of war operations (Arkin 2009), which as explained above is a particularly relevant research field within machine ethics and roboethics. The external module proposed in this paper would incorporate additional characteristics to Arkin's moral governor, such as evolving to incorporate any updates, exceptions, or patches that might be deemed necessary given the ever-evolving experience, not only with a particular device but also with all devices belonging to the same class. Another difference is that Arkin proposes that there can be overriding on the ethical control, albeit "with upmost certainty on the part of the operator". The external module in this paper could contemplate overriding as a "big-red-button" decision, but only before the external module, so that it should not be possible to override the top-down rules in any context. The behavior of the device should be constrained by the external module regardless of the circumstances. The external module must be separate from the machine so that it can remain unaffected by the specific users or programmers of the autonomous device. This requires all autonomous devices to be connected to the external module (e.g., via Wi-Fi) to function and calls for immediate pause of all actions in the event that this connection is lost.

It should be noted that building such an external module is plausible in principle, but it would be by no means an easy task; implementing such a system entails two main risks:

> Defining the set of agreed rules in an objective and transparent way, as well as the possible outcomes of any action belonging to the space of decisions.
>
> Establishing the consequences of actions in complex or uncertain settings. One example tool that could be applied for this is inference engines, which would be able to test whether any undesired consequence arises from the knowledge base of the device or the intended action. In addition, fuzzy logic can be used to deal with the degree of ambiguity that is usually present in real-world decisions.

The insurance problem manifests a *need for transparency* in the decision making of automata. Given that the strategy developed by the robot should ultimately be understandable by human beings, not all algorithms should be valid to develop automata, regardless of their good performance. For instance, approaches based on artificial neural networks should be examined carefully, as they commonly lead to black-box solutions where the logic underlying the decision is not easily identifiable. On the contrary, approaches such as size-constrained decision trees would be extremely useful, as they can, in small cases, be understandable in terms of human behavior. Along the same lines, knowledge extraction techniques can effectively clarify the results of complex optimized algorithms. These methods should be used extensively as the need for transparency grows larger.

Last, even if knowledge extraction is not practical in a particular instance, law-enforcement can rely on test cases (in the example, feed several profiles to the algorithm and assess whether decisions are discriminatory up to a given percentage). That is, if it is not possible to establish a priori if the rule is followed, it can always be tested ex-post. This last-resort option should be avoided if possible.

With these two proposals in mind (filtered decision making and a restriction on suitable algorithms for the sake of transparency), it should be reasonable to establish behavior codes to be abided by self-evolving machines in a safe and transparent way. These codes should be updated regularly as more information is known about the possible courses of action for the automata. Filtered decision making would avoid scenarios where machines get out of control.

5. The Appeal of Positive Ethics and Subjectivity as a Hard Limit

The approaches based on top-down, negative ethics seem to be far from the complete experience of morality in human beings. The rationale of decision making by seeking the best consequence in terms of a given objective (such as a common good or maximizing a given definition of utility) is undoubtedly appealing.

These utilitarian or consequentialist approaches have been criticized in the past because of the computational problems that might arise when calculating the consequences of any given action. Calculating first-order effects might be difficult, but even in the simplest of circumstances, calculating all cascade of effects quickly becomes an unmanageable task.

However, there is a more problematic issue underlying utilitarian approaches: the definition of the utility function is not always obvious. In order to be implementable, it must be expressed in the programming language the machine is built on. This might not seem like a restrictive condition: it is easy to build a trading automata dealing in a stock exchange to maximize profit, or profit for a given risk. The utility function that guides the trading automaton is purely beneficial, which can be expressed as the amount of money made in the trades it closes in the market. Similarly, a robot with a higher human interaction such as a robotic hotel receptionist might have its utility function defined as being agreeable to customers. In order to give a numerical value to this, we might ask clients to fill a quick survey about their satisfaction. The receptionist robot will try different strategies and select the ones that please customers the most as revealed in the surveys.

The past examples are easily understandable and immediately implementable. However, the proponents of bottom-up machine ethics go far beyond this conception. For instance, Kurzweil as a representative of the most extreme versions of technooptimism, believes that machines will not only behave in a way consistent with values, but experience these values themselves and become, in his words, *spiritual machines* where, from an artificial intelligence basis, consciousness and authentic ethics will *emerge* (Kurzweil 2012).

There have been very compelling proposals dealing with bottom-up values in other sciences, particularly with respect to the plasticity of values, which are understood as inherently subjective but built in cooperation with other human beings. Some authors, such as E.O. Wilson, have long ago discussed the emergence of values in societies (Wilson 1975). Interestingly, values such as cooperation for a common good have been explained in an evolutionary framework as a survival-optimizing strategy, as in the works of Danielson, (Danielson 1998). Evolutionary game theory (EGT) is extremely interesting in this context. However, I would argue that the behaviors that emerge in EGT and that we categorize as "ethical" are not guided by moral values themselves (such as helping the weak) but are rather the result of adapting to a very different goal (survival of the individual, the family or the community). Some of the strategies that emerge seem to have a moral component to it (such as helping the weak) and result in eusocial behavior in general, while some others exemplify selfish behavior (e.g., cuckoo-like strategies). I would argue that these behaviors are guided by survival and not by any moral values, even if a human observer could interpret them as being moral or immoral depending on the case. This categorization might be interesting and useful, but giving a useful description should not be confused with guiding behavior itself.

The thinkers that defend emergence in machines stay silent about the possible mechanisms that can give origin to this phenomenon. For instance, Kurzweil states that a complete, functioning mind can be created by adding a *value module* to a brain based on pattern recognition (Kurzweil 2012), but does not give any details on how this module should be created or how it can be coded. All these bottom-up proposals have in common this lack of theoretical support at this point. It should be noted that the three concepts are usually understood to be related: values emerge within consciousness, and consciousness is one of the main properties of subjectivity. The emergence of values presupposes the emergence of subjectivity and consciousness. There have been some attempts at studying the characteristics of current AI in relation to the properties of consciousness, such as the interesting work

of Kuipers (Kuipers 2008, pp. 155–70). However, these descriptive studies do not provide any proof of the existence of subjectivity in the machines.

Very importantly, the emergence of values should not be confused with the emergence of a function to fit externally defined values. An objective function defining a value can effectively emerge (e.g., via reinforcement learning, where the output of a given function could be adjusted to the value given by a user – for instance, evaluating whether a decision is fair or not). In this case, *the function to evaluate fairness* emerges, rather than the *concept of fairness itself*. This distinction is very important, as fitting a function does not require any subjective experience and can effectively be accomplished—even if the task at hand is difficult. This would be an example of possible "positive ethics", as an automaton can be designed to maximize fairness as long as it is given an input including this evaluation (in this case, this could be expressed as a fairness evaluation coming from an external judge. This fairness evaluation would be used as the desired output for supervised learning). Once again, in this case, a fitness function emerges, but the concept of justice comes from the outside in the form of the outputs for supervised learning. However, because the concept of fairness, rather than emerging within the learning automaton, is received by it, this does not constitute an example of bottom-up ethics. The strategy to optimize the objective function emerges bottom-up, but the objective function itself comes from the outside. This is discussed further in the next section.

6. Language as a Limit

It has been said that living beings are the only *beings with a purpose* (Rosenbrock 1990). However, machines can undoubtedly have a purpose as well, as long as it is given to them by the programmer. This purpose is, in the case of machines, the objective function that is optimized[4] by the learning algorithm implemented in their software. Utilitarian approaches to machine ethics take this as their starting point.

There are two elements that seem to be missing from the public debate. First, in order to be programmed, the objective function that will guide the machine's actions needs to be expressed in the same programming language used by the machine. By nature, this programming language must be a *public* one in the sense of being understandable by all the users of the language in the same way (Leach 2011). It is useful to remember the distinction proposed by Javier Leach between public languages (objective and understood by all users in the same way) and *personal* ones, which include all the concepts in the public ones and expand them to include subjective information and experience. Mathematics or programming code would be examples of public languages, while concepts referring to emotions (such as anger or disgust) or values (such as justice or beauty) would clearly belong to the personal sphere. It should be noted that the degree of subjectivity of a term varies widely, with "chair" being less exposed to subjectivity than "non-combatant" in war operations terminology, and values such as "fairness" being arguably the most subjective of them all.

It should be understood that the public domain includes only completely objective concepts, qualities, and magnitudes that can be impartially determined or measured. The public domain is the realm of the sciences and of programming. "Length", "weight" or "wavelength" belong to thepublic domain.

Any personal concept can be projected into the public domain at the expense of losing some of its nuances. If the degree of subjectivity is small, the projection will be relatively faithful to the personal concept. This would be the case if we project "color", for instance, as a table from wavelengths, which can be measured objectively, to words such as "red" or "blue". If the automaton has a very specific purpose and a narrow set of possible interactions with human beings, these projections can be defined in a relatively straightforward way. For instance, we could project the concept "avoiding harm", for a

---

[4] I include, within the broad term "optimization", approaches such as constraint satisfaction, which can be used to describe any general decision problem.

self-driving car, as "avoiding collision with another car or a human being". However, this should not distract from the fact that the value has not been represented in full.

If we project "distributive fairness" into "distributing wealth into equal portions" (assuming we have previously defined terms such as "wealth" and "portions"), we might incur a larger error and will probably have considerable trouble agreeing on a definition. While we can build attempts to express a concept from the personal languages in a public way, this attempt can never capture the true essence of the concept.

We can train a robot that distributes food to calculate equal portions by measuring the weight of the trays it serves. However, that is far from understanding the concept of justice. In order to be able to program a given purpose in an automaton, this purpose must be amenable to expression in the programming language. Surprisingly, these considerations are missing from the current discussion, which seems to disregard the different types of language. It is dangerous to assume that the two spheres of language (public and personal) are the same, because it gives us the false notion that robots can act ethically by their own means and have authentic moral judgments, when they can only operate based on approximations expressed in a public language. In addition, this takes away the focus from the imminent task of the programmer: providing an objective function that is as similar as possible to the concept, which is not expressible in those terms. Happiness and justice belong to the personal sphere, so they cannot be used to define an automaton's purpose. However, profit and loss can be easily expressed as a number, so it can indeed be used to guide a robot's actions.

Nothing suggests that automata based on current technology will ever be able to deal with personal language (which seems to be linked to subjectivity) or have a subjective experience of their purpose—at the very least, there is nothing to make us think otherwise. Some authors seem to believe that there will be some emergence of subjectivity that could explain how automata would be able to deal with values and spirituality. The proponents of an emergence of subjectivity, as expressed in the previous section, have not been able to provide any proof or clues on what might underlie this phenomenon. Their purpose is not in themselves, but is rather given to them by the programmer; it comes from the outside. This means that this purpose will always be defined in terms of their relationship to human beings or the environment.

The responsibility of defining this objective function will always be the programmer's. In addition, assigning a good purpose definition is not enough. In the example of the robotic hotel receptionist, the purpose of the automaton was clear: improve client surveys. The robot could learn strategies to make clients more satisfied by trial and error. However, it could also learn to tamper with survey data, which could also lead to better results, albeit in a non-desired way. This sort of behavior can happen because the expression of value in that case—i.e., client satisfaction—in a public language—i.e., survey numbers—is not perfect. This example also shows the need for negative ethics as proposed in Section 4. Any learning, self-evolving robot should be subject to filtered decision making, where any unacceptable action should be detected and prevented before it happens. What is more, this example shows the importance of building the filtering module online: a priori, the tampering behavior probably was not anticipated. Once it has happened, however, a new rule can be added to the code abided by receptionist machines. This could be made effective immediately to prevent any future instances of the problem.

In short, negative approaches will always be necessary, while positive ethics can be implemented in some instances. The basic requirement for the latter is expression in a public language. This definition will be carried out by the programmer and will correspond to a projection of values into the public language, and should not be confused with the bottom-up emergence of values themselves.

The automaton can then develop, bottom-up, a strategy to optimize an objective function, but it is not able to learn or derive values themselves, which belong to the personal sphere and which the machine does not comprehend, as it only grasps approximate depictions expressed in a public language—the only one it understands. This rules out the possibility of spiritual machines that develop their own transcendent values, at least with current technology based on public languages. In addition,

this stresses the responsibility of the programmer when creating the value approximations that will guide the machine's actions.

7. Conclusions

The growing field of machine ethics has proposed several strategies for endowing machines with ethical behavior. This paper has presented a classification that organizes these proposals around a double dichotomy. Negative ethics is concerned with preventing harm to other beings, while positive ethics focuses on creating the largest good. Top-down ethics conceive the moral rules or the definition of ethical good as something external that is adopted by the subject, which contrasts with bottom-up ethics, which considers that it is the subject who builds the values that guide her actions. All approaches to machine ethics can be classified using these categories.

Within negative ethics, ethics based on moral duty are difficult to implement. Law-based approaches have the advantage of being more easily implementable as they are objective. This paper proposes *filtered decision-making* as a framework, where the decisions of the automata should be approved by an external objective module. This filter checks that the robot's decision complies with the set of rules it is ascribed to and, in the event that the decision is noncompliant, prevents it from being enacted. This approach also has the benefit of providing an easy way of updating the legal codes for all automata at the same time.

Within the context of positive ethics, some authors have envisioned robots that evolve their own sense of moral consciousness in a bottom-up manner. However, they do not provide any ideas for explaining how this could happen. This paper has emphasized that current technology is based on public languages, so that values can only be projected from the public sphere in an approximate way. Machines can take their objective function from the programmer and develop bottom-up strategies to optimize it. However, they cannot emerge, bottom up, this objective function or the values that originated it. This discards the possibility of spiritual machines that develop their own transcendent values, at least with current technology. Language and subjectivity are, therefore, a limit for machine ethics.

Although machines cannot have a purpose in themselves, they can indeed have *a purpose*, as long as it is given to them by the programmer (in a top-down instance of ethics), so it would be a purpose *outside of themselves*. This must be expressed in a public language—which means, in other words, that it must be amenable to be written in a programming code. In addition, given that evolving machines can have unexpected behaviors, additional negative ethics in the form of filtered decision-making is still necessary. Only by combining a law-based approach with a well-understood definition of purpose (outside themselves and expressed through a public language)—that is, understanding the limits of present-day machine ethics—will we be able to tackle the challenges that lay in the future field of robotics.

Conflicts of Interest: The author declares no conflict of interest.

References

Allen, Colin, Iva Smit, and Wendell Wallach. 2005. Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology* 7: 149–55. [CrossRef]

Arkin, Ronald. 2009. *Governing Lethal Behavior in Autonomous Robots*. Boca Raton: CRC Press.

Asimov, Isaac. 1950. *I, Robot*. New York: Gnome Press.

Avraham, Ronen, Kyle D. Logue, and Daniel Schwarcz. 2012. Understanding Insurance Anti-Discrimination Laws. Available online: http://repository.law.umich.edu/cgi/viewcontent.cgi?article=1163&context=law_econ_current (accessed on 19 May 2017).

Campbell, Robert L., John Chambers Christopher, and Mark H. Bickhard. 2002. Self and values: An interactivist foundation for moral development. *Theory & Psychology* 12: 795–823.

Casey, Bryan James. 2017. Amoral machines, or: How roboticists can learn to stop worrying and love the law. Available online: https://ssrn.com/abstract=2923040 (accessed on 1 May 2017).

Danielson, Peter. 1998. Modeling Rationality, Morality, and Evolution. Oxford: Oxford University Press on Demand.
Floridi, Luciano. 2005. Information ethics, its nature and scope. *ACM SIGCAS Computers and Society* 35: 3. [CrossRef]
Greene, J.D., R.B. Sommerville, L.E. Nystrom, J.M. Darley, and J.D. Cohen. 2001. An fMRI investigation of emotional engagement in moral judgment. *Science Magazine* 293: 2105–8. [CrossRef] [PubMed]
Handelsman, Mitchell M., Samuel Knapp, and Michael C. Gottlieb. 2009. Positive ethics: Themes and variations. In *Oxford Handbook of Positive Psychology*. Oxford: Oxford University Press, pp. 105–13.
Hardy, Sam A., and Gustavo Carlo. 2011. Moral identity: What is it, how does it develop, and is it linked to moral action? *Child Development Perspectives* 5: 212–18. [CrossRef]
Head, Simon. 2014. Mindless: Why Smarter Machines are Making Dumber Humans. New York: Basic Books.
Honderich, Ted. 2005. The Oxford Companion to Philosophy. Oxford: Oxford University Press.
Howard, Ronald Arthur, and Clinton D. Korver. 2008. Ethics for the Real World: Creating a Personal Code to Guide Decisions in Work and Life. Cambridge: Harvard Business Press.
International Federation of Robotics (IFR). 2016. *World Robotics 2016*. Frankfurt: International Federation of Robotics.
Jackson, Frank. 1991. Decision-theoretic consequentialism and the nearest and dearest objection. *Ethics* 101: 461–82. [CrossRef]
Kahneman, Daniel. 2011. Thinking, Fast and Slow. New York: Macmillan.
Kant, Immanuel, and Thomas Kingsmill Abbott. 2004. Critique of Practical Reason. Miami: Courier Corporation.
Kuipers, Benjamin. 2008. Drinking from the firehose of experience. *Artificial Intelligence in Medicine* 44: 155–70. [CrossRef] [PubMed]
Kurzweil, Ray. 2012. How to Create a Mind: The Secret of Human Thought Revealed. London: Penguin.
Lapsley, Daniel K., and Darcia Narvaez. 2006. Character education. In *Handbook of Child Psychology*. New York: John Wiley & Sons.
Leach, Javier. 2011. Mathematics and Religion: Our Languages of Sign and Symbol. West Conshohocken Templeton Foundation Press.
Lichtenberg, Judith. 2010. Negative duties, positive duties, and the "new harms". *Ethics* 120: 557–78. [CrossRef]
Luxton, David D. 2014. Recommendations for the ethical use and design of artificial intelligent care providers. *Artificial Intelligence in Medicine* 62: 1–10. [CrossRef] [PubMed]
Martin, Jack. 2004. Self-regulated learning, social cognitive theory, and agency. *Educational Psychologist* 39: 135–45. [CrossRef]
Powers, Thomas M. 2006. Prospects for a kantian machine. *IEEE Intelligent Systems* 21: 46–51. [CrossRef]
Rosenbrock, Howard H. 1990. *Machines with a Purpose*. Oxford: Oxford University Press.
Slote, Michael A. 1985. *Common-Sense Morality and Consequentialism*. Abingdon-on-Thames: Routledge & Kegan.
Van de Voort, Marlies, Wolter Pieters, and Luca Consoli. 2015. Refining the ethics of computer-made decisions: A classification of moral mediation by ubiquitous machines. *Ethics and Information Technology* 17: 41–56. [CrossRef]
Veruggio, Gianmarco, Fiorella Operto, and George Bekey. 2016. Roboethics: Social and ethical implications. In *Springer Handbook of Robotics*. Berlin and Heidelberg: Springer, pp. 2135–60.
Wilson, Edward O. 1975. Sociology: New Synthesis. Cambridge: Belknap Press.
Yampolskiy, Roman, and Joshua Fox. 2013. Safety engineering for artificial general intelligence. *Topoi* 32: 217–26. [CrossRef]