The Springer Tracts in Advanced Robotics (STAR) publish new developments and advances in the fields of robotics research, rapidly and informally but with a high quality. The intent is to cover all the technical contents, applications, and multidisciplinary aspects of robotics, embedded in the fields of Mechanical Engineering, Computer Science, Electrical Engineering, Mechatronics, Control, and Life Sciences, as well as the methodologies behind them. Within the scope of the series are monographs, lecture notes, selected contributions from specialized conferences and workshops, as well as selected PhD theses.

Special offer: For all clients with a print standing order we offer free access to the electronic volumes of the Series published in the current year.

Indexed by DBLP, Compendex, EI-Compendex, SCOPUS, Zentralblatt Math, Ulrich's, MathSciNet, Current Mathematical Publications, Mathematical Reviews, MetaPress and Springerlink.

More information about this series at http://www.springer.com/series/5208

Jean-Paul Laumond · Emmanuelle Danblon · Céline Pieters
Editors

Wording Robotics

Discourses and Representations on Robotics

Springer

Editors
Jean-Paul Laumond
LAAS-CNRS
Toulouse, France

Emmanuelle Danblon
ULB, Service de Linguistique
Brussels, Belgium

Céline Pieters
LAAS-CNRS
Toulouse, France

ISSN 1610-7438 ISSN 1610-742X (electronic)
Springer Tracts in Advanced Robotics
ISBN 978-3-030-17973-1 ISBN 978-3-030-17974-8 (eBook)
https://doi.org/10.1007/978-3-030-17974-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Foreword

Robotics is undergoing a major transformation in scope and dimension. From a largely dominant industrial focus, robotics is rapidly expanding into human environments and vigorously engaged in its new challenges. Interacting with, assisting, serving, and exploring with humans, the emerging robots will increasingly touch people and their lives.

Beyond its impact on physical robots, the body of knowledge robotics has produced is revealing a much wider range of applications reaching across diverse research areas and scientific disciplines, such as biomechanics, haptics, neurosciences, virtual simulation, animation, surgery, and sensor networks among others. In return, the challenges of the new emerging areas are proving an abundant source of stimulation and insights for the field of robotics. It is indeed at the intersection of disciplines that the most striking advances happen.

The *Springer Tracts in Advanced Robotics (STAR)* is devoted to bringing to the research community the latest advances in the robotics field on the basis of their significance and quality. Through a wide and timely dissemination of critical research developments in robotics, our objective with this series is to promote more exchanges and collaborations among the researchers in the community and contribute to further advancements in this rapidly growing field.

In this book, Emmanuelle Danblon, Jean-Paul Laumond, and Céline Pieters present a collection of rare perspectives on the relationship between human language and the common perception of robotics. The work is based on a unique multidisciplinary gathering organized in the framework of Actanthrope, a European Commission's project. Exploring the role human language has on public understanding of robotics, *Wording and Robotics* offers remarkable new insights collected from diverse point of views of a pluri-disciplinary group of researchers and scientists in the fields of robotics, neurophysiology, rhetoric, and philosophy.

Artificial Moral Agents. Really?

Mark Hunyadi

Abstract How can we plausibly refer to robots as artificial *moral* agents? Considering the useful classification of the philosopher of the field of artificial intelligence James H. Moor, who identified four different kinds of ethical, I will argue that the term of artificial *moral* agent is philosophically illegitimate. My argumentation is developed in three stages: the first stage addresses the actual choice of the ethical principles to be programmed into the machine; the second stage explores the difficulties inherent in giving these principles an algorithmic form; and the third focuses on the supreme difficulty arising from the very nature of moral reasoning. This analysis aims at encouraging the research on the concepts of moral reasoning and judgement. Indeed, a fine understanding of these notions should reveal the full extent of the problem with artificial moral agents; before we can discuss *machine ethics* or artificial ethics, we must, if we are to avoid speculation and ideology, have a clear understanding of what ethics is, what type of rationality it implements, and what is the nature of ethics and ethical conduct in general.

Keywords Moral · Artificial · Agent · Ethics · Moore

1 Explicit Ethical Agents

It has long been true that tools and machines generally have an *ethical impact*. Firstly because, from a superficial perspective, tools are an extension of human ability and can therefore be used for both worthwhile and reprehensible purposes. Secondly because, from a deeper perspective, as soon as they are invented and come into widespread use, they change human behaviour by creating new behavioural expectations. For example, use of the watch imposed an expectation of punctuality

Translated from French by Katherine Mérignac.

M. Hunyadi (✉)
Université Catholique de Louvain, 1 Place de l'Université, 1348, Ottignies-Louvain-la-Neuve, Belgium
e-mail: mark.hunyadi@uclouvain.be

and greater discipline in the workplace, mobile telephones have brought a growing expectation of availability, and so on.

Today, however, we are no longer talking about the ethical impact of machines, but about artificial moral agents (AMAs), in other words machines capable of making decisions under uncertainty. What I would like to explore today is in what sense we can plausibly refer to these machines as artificial *moral* agents. Allow me in this regard to borrow a recent and extremely useful classification from the philosopher of artificial intelligence James H. Moor who, in an article published in 2006, identified four different kinds of ethical robot [1, 2]:

1. **'Ethical impact agents'**—agents that have an ethical impact. This simply refers to the ethical impact that robots or any other tool or machine can have on human beings, in the sense mentioned above.
2. **'Implicit ethical agents'**—these are machines that perform a specific technical function but are designed in such a way as to prevent ethically undesirable consequences. They have a built-in 'ethical component' to prevent adverse effects. Moor mentions user safety considerations for example, but one could also mention considerations relating to discrimination in hiring, renting, lending and privacy protection [3],[1] or military robots that distinguish between civilians and non-civilians. We refer to them as *implicit* agents because the tasks that they perform are not in themselves ethical (they are, broadly speaking, technical tasks), and because the 'ethical component' that restricts their freedom of action is invisible.
3. **'Explicit ethical agents'**—these are machines designed specifically to take decisions of a so-called ethical nature. They are programmed to resolve 'ethical' problems in line with explicit moral principles. The field of application most frequently cited for these agents is medical ethics. Basically speaking, the question that these explicit ethical agents must answer is this: in a moral dilemma, given the information available, what is the best course of action from an ethical perspective? The subtitle of an important book on machine ethics lends additional pathos to this question: *Teaching Robots Right from Wrong*.
4. **'Full ethical agents'**—these machines are allegedly endowed with a code of ethics in the human sense, and therefore have the ability to take initiatives and make judgements about human actions in general. So, from an ethical perspective, they are (arguably) no longer distinguishable from human beings. In that case they should not, strictly speaking, be referred to as *agents*, but as *actors*.

The literature today seems, by consensus, to focus the issue of so-called artificial moral agents on the third type of machine: explicit ethical agents. The fourth type is still eminently speculative, and the first two seem to include all human artefacts rather than robots specifically. The third type, on the other hand, raises some very interesting

[1] Wallach and Allen, who do not make these distinctions in their book, are nevertheless clearly thinking about this when they say: "Moral agents monitor and regulate their behaviour in light of the harms their actions may cause or the duties they may neglect. Humans should expect nothing less of AMAs. A good moral agent is one that can detect the possibility of harm or neglect of duty, and can take steps to avoid or minimize such undesirable outcomes" [2].

questions, since we are talking about machines being able to make ethical decisions in problematic situations. So the guiding question here is: how far can artificial agents be considered *moral* agents as suggested by the acronym AMA, which, in a manner of speaking, lends credence to the possibility and gives it an almost familiar feel? I will argue that, even in the narrow sense of the third type of machine, the term artificial moral agent is philosophically illegitimate.

I will develop my argument in three stages: the first stage addresses the actual choice of the ethical principles to be programmed into the machine; the second stage explores the difficulties inherent in giving these principles an algorithmic form; and the third focuses on the supreme difficulty arising from the very nature of moral reasoning.

2 Stage One: Moral Agents Versus Executors

First, let's look simply at how the term 'artificial moral agents' is understood when it refers to explicit moral agents. According to Anderson and Anderson [4], the purpose of these robots is "to do the morally correct action in an ethical dilemma and justify it." And they add: "All that is required is that the machine act in a way that conforms with what would be considered to be the morally correct action in that situation and be able to justify its action by citing an acceptable ethical principle that it is following."

First, it should be pointed out that, from this perspective, *machine ethics* is a relatively modest goal: it involves successfully identifying an action which, in an indeterminate moral situation, would be regarded by human beings as meeting current ethical standards. What is modest about this is, among other things, that the machine is not designed to be morally *responsible*, therefore capable of free initiative. It is designed simply to calculate the most ethical solution considering a set of previously implemented principles and a pre-defined context—for example, the medical sector. Let me re-phrase that: in this approach to *machine ethics*, we are seeking an *outcome* (in this case an action) produced by an algorithm that has successfully coded a moral principle. The aim is to achieve a specific outcome using a specific algorithm. As I understand it, that is the definition of an explicit moral agent, which plays a key role in machine ethics research today.

So, let's stay with this approach to *machine ethics* for now. From the roboticist's perspective, two interconnected but very different issues arise here. The first is knowing *which rules* to encode as algorithms, which raises the philosophical question of *which ethics to choose*. In other words, 'what' is the most relevant choice of principles. The second issue is knowing how to *program* decision-making procedures, which raises the technical question of 'how' to implement these principles in the system. In this respect, there are several possible models: *machine learning*, *deep learning*, *neural network approach*, etc. I will be addressing only the first of these two problems, as the second cannot be resolved without resolving the first. And it is this that I will focus on doing.

When we think about designing artificial moral agents, the question of which ethics to choose inevitably arises. It's a huge problem! There are already numerous types of ethics in our social relationships. The *same* society can therefore be utilitarian in its approach to immigration, deontological when it comes to voting rights, meritocratic in the distribution of rewards, etc. However, even within a narrow area like medical ethics (which roboticists regard as the ideal area of application for *machine ethics*), there are several competing ethical paradigms; and even within these paradigms, there are conflicting norms, as every doctor knows (for example, the four canonical norms: beneficence, non-maleficence, autonomy and justice). In this respect, there is a fundamental ethical indeterminacy inherent to ethics themselves, or to ethical judgement. As far as *machine ethics* is concerned, this means one thing: if you program a specific set of ethical principles into a machine, you do not make the machine an artificial *moral* agent, but an *executor of those specific principles*, which is an entirely different thing. This so-called 'artificial agent' will be expected to respond according to *those ethical principles,* chosen by the programmer. Strictly speaking, then, the correct term to use is not artificial moral *agent*, but *executor*; and not artificial *moral* executor, but artificial *utilitarian, deontological* or *perfectionist* executor, depending on the ethical principles chosen by the programmer. Therefore, 'explicit moral agents' are not, strictly speaking, either agents or moral agents.

The choice made by the programmer is necessarily to the detriment of all other available possibilities; being arbitrary, it is a considerable, artificial and undue reduction from the outset. The scale of this initial reduction is generally trivialised in the literature, either because it is considered simply as inevitable, or because it is taken for granted and given little importance. On the contrary, I believe it is very important, and I will show again later exactly how it undermines the nature of moral reasoning. But another highly problematic consequence of this choice—which is *not related* to the nature of moral reasoning—is the *concentration of power* in the hands of programmers, who therefore dictate the ethicality of machines and hence of the human behaviour induced by these machines. Yet given that robots such as these will increasingly occupy our social environment in the future, this is a major social, political and anthropological concern. I will not be elaborating on these general consequences here, however—even though, in my view, they are of the utmost importance.

That said, I have already demonstrated why explicit moral agents should not be referred to as either agents or moral. This was, so to speak, the first stage in my reasoning, which is based on the argument that the mere fact of having to program a machine violates our most immediate moral experience, deriving from the plurality of available moral grammars. Basically, this simple demonstration should be enough to seriously temper our ambitions for machine ethics.

However, I would now like to put forward some more fundamental reasons, connected with the very nature of moral reasoning, why it is generally and, strictly speaking, impossible at present to use the term 'artificial *moral* agents'. What I want to show over the following two stages are the challenges that the specific nature of moral activity presents to researchers in artificial ethics. *If* artificial moral agents are to be at all possible, then in my view researchers must address the difficulties specific to moral activity itself.

3 Stage Two: The Ethical Algorithm (Three Difficulties)

The arguments I am going to put forward in this second stage are therefore based on a fundamental philosophical reflection on the nature of ethics. And this allows me to make a preliminary remark: after skimming through the scientific literature on AMAs, I am amazed by the cavalier approach to ethics themselves. Even though we hope to create ethical machines and are therefore under the technical obligation to artificially recreate our own natural ethics step by step, we rely on vague notions about the nature of ethics, we employ concepts such as obligation, good and virtue almost at random, and make do with sweeping statements about accountability and very loose convictions about the nature of moral judgement and reasoning. Extreme specialisation in artificial intelligence goes hand in hand with extreme moral amateurism, which is surprising and, in my view, can only hinder research into AMAs in the long term. Admittedly, the authors usually call for cross-disciplinary cooperation with philosophers but (to the best of my knowledge of course) there is no real evidence that this is anything but wishful thinking.

The fundamental question that concerns me now is not which ethical principles should be programmed into machines, but whether ethics can, generally speaking, be programmed or computed. Once a guiding ethical principle has been selected, can it be given an algorithmic form? This is an unavoidable question for researchers in the artificial intelligence field: if machine ethics are to be possible, then they must be formalisable. Are they? To even consider developing an artificial agent of such and such an ethical principle, roboticists must assume that they are. They must be capable of giving moral reasoning an *algorithmic form*, of making it computable. This is, in fact, a decisive stage in machine ethics and in artificial intelligence on the whole; a stage on which everything else depends.

What I would like to demonstrate is that because machine ethics raises this issue in this manner, i.e. in terms of computability—and because it *has no choice* but to raise it in this manner—it makes the issue itself impossible to resolve. The point is that, for reasons relating to the very purpose of algorithms, and to what algorithms are, machine ethics must focus on the desired *result*, or *outcome*, in this case an action that we hope will be the best possible option. In a complex situation where ethical principles are at stake (never mind which for now), we want to know the right thing to do, in other words the right decision to take. Consequently, the ethics of AMAs could be described as 'outcome-oriented ethics', which, in the philosophical tradition, are also called *normative ethics* and are directed towards problem solving. The question that this type of ethics must generally address is the following: based on what we know and on the principles that we want to apply, what is the best thing to do?

In this respect, I would say: if ethics really did come down to straightforward problem solving (like in a game of chess or Go), I would see no principled reason why designing AMAs should not be possible. The resolution of problems would, after all, just depend on their complexity, the specific circumstances, the availability of relevant information, etc., as with any technical problem. However, the problem

is that the prior decision to consider ethics as being 'outcome-oriented', which is an inevitable decision given the current state of play in computing research, is a decision that, in itself, prevents all means of tackling the problem it is supposed to resolve. This approach, which is inherent in computation, is based on a positivist reduction that is incompatible with the nature of moral reasoning. Let me explain.

(1) This reduction is in fact already revealed by the use of the word 'agent' in the abbreviation AMA. An agent is someone or something that *acts*, irrespective of whether it is capable of thought or not; it is therefore someone or something that *produces an action*, in other words an event that influences the course of the world. As we want to program a robot to *do something*, there has to be a kind of 'fetish for action' inherent in the robotic procedure itself, since the whole point is to obtain a certain action.[2]
However, the first point I would like to make is this: while it is true that ethics is naturally geared towards action, which is why it comes under a branch of philosophy called practical philosophy, action is far from being all there is to ethics. If it were, ethics would simply be a behavioural science and the only problem would be achieving behaviour consistent with what is deemed 'good'.[3] In other words, although action is to some extent the end point of ethics, it is nonetheless just the visible, emerged and positively identifiable part, the tangible manifestation of an invisible process. Moreover, there are many actions that *do not* result from an ethical process, for example reflexes, impulses, routine behaviour, role playing, obeying an order or fulfilling a request. What gives an action-oriented process its *morality* is the 'grounds' for the action. Therefore, it is not the action in its materiality that makes the difference, but the whole *process* leading up to the decision to act in a certain way. If an action is ethical, it must always be *justifiable* from a *moral standpoint*. So a moral process always consists of a visible space of action and, leading up to that action, an invisible space of *reasons*. These are the reasons that govern and justify the action, should justification be required. And who should be able to explain these reasons? The actor of course, the person taking the action. Since the space of reasons is invisible to outside observers, only the actor can explain the reasons for his or her action.

(2) This remark on moral epistemology—that only the actor can truly explain his or her action—is very important in our context and brings me to my second point. I have already quoted Anderson and Anderson, according to whom: "All that is required is that the machine act in a way that conforms with what would be considered to be the morally correct action in that situation and *be able to justify its action by citing an acceptable ethical principle that it is following*" (emphasis added). They argue that the second requirement for ethical machines, which I have not discussed yet, is that they be able to *justify* their action by *citing* the ethical principle that they are following. Thus, these robotics philosophers clearly recognise that not only the action itself, but also its justification—the invisible part of the action—is an essential component of ethical conduct since they require it in their machines.
But in what sense can one speak of *justification* here? As I have just said, justification is something that actors themselves must provide, in the first person, as a reason for their actions. For a machine, however, the reason can only be the algorithm that it has followed, and which it was programmed to follow in a given type of situation. The justification for its action is no more than an *update* of its mechanism. In these circumstances, justification amounts to *explaining* a decision (the underlying principle of which could be formulated identically by the *external* programmer), or to revealing the *causes* of a decision rather than *justifying* it, that is to say providing the *reasons* behind it. In other words, the requirement for justification, as described by Anderson and Anderson [4], is aimed at *clarifying* the machine's decision, but does not by any means guarantee its *morality;* enunciating a computing sequence is not the same thing as choosing a reason, as a moral agent would, it is just making an invisible mechanism visible.

(3) Now, even discounting the fact that such a justification is not a justification at all, but merely an update of calculation mechanisms, there would still be a major problem with the calculation itself. This, for me, is the third difficulty associated with the use of algorithms. I call it *the problem of the informational basis* of the ethical algorithm: to correctly judge a situation from a moral perspective, what inputs would the machine need? What informational basis would be required to carry out an adequate moral evaluation of a situation? If, for example, we took a liberal view and developed a moral algorithm that, as a rule, extended individual liberties, how would we know if an action increased or restricted those liberties? How would we measure that?
To deal with this difficulty, AI theorists automatically turn to the ethical theory that most resembles computation: utilitarianism. Even so, things are far from clear. As we well know, there are very many versions of utilitarianism, precisely because we do not really know what to compute; in other words, what should be taken into consideration when computing well-being. Herein lies the problem with the informational base. This difficulty is inherent to the utilitarian theory of ethics despite the fact that it is based on computation; it would be a thousand times greater in the case of qualitative ethical theories that focus on happiness, fulfilment and self-realisation, all of which are equally valid yardsticks for defining our moral attitudes

[2]This fetish for action is emblemised by the canonical example of the trolley problem, which is regarded as 'fundamental' to moral reasoning. This example, first presented by Philippa Foot in 1967, has since been regarded as the touchstone of ethics. And it does indeed work well with machine ethics; unsurprisingly, in [2] it is used to introduce the general theme.

[3]The distinction between conduct that conforms to ethics and ethical conduct itself is exactly what Kant had in mind when he distinguished between acting in accordance with duty and acting out of duty.

4 Stage Three: Ethical Counterfactuality

At the second stage of this discussion on AMAs, I talked about the difficulties connected with the algorithmic approach: first, it places undue emphasis on action; second, it misrepresents justification; and third, there is the problem of the informational basis.

The supreme difficulty is however encountered at the third stage and relates to the nature of moral reasoning itself. The difficulty is this: *the moral principle by which a given situation should be judged is never self-evident.* It is never obvious how best to approach a moral situation. The 'grounds' on which we morally judge a situation in which we would like to intervene do not derive from the situation itself, but from the mind of the person judging it. Who can say, for example, whether our approach to abortion should be guided by the basic rights of the mother and the unborn child, the greater good theory propounded by utilitarians, virtue ethics, pro-life values or religious beliefs; or whether any of these moral grammars should be balanced against other economic, cultural or legal considerations and, if so, to what extent? Of course, the answer will not come from the situation itself but from those called upon to decide and act accordingly, depending on how they perceive and interpret the situation. The situation will not interpret itself. What makes an abortion situation ethically problematic is that it can be interpreted in several different ways. If there were only one possible interpretation, it would not be an ethical problem but a technical one.

If situations do not interpret themselves, it is because they fall within the realm of the factual; on the other hand, the principles by which we judge them fall within the realm of the *counterfactual*. The situation does not tell us which principles to apply, it is the actor who decides to apply such and such a principle to the situation. Yet all principles are counterfactual: they represent an *ideal* reality capable of guiding an agent's judgements and actions. Counterfactuality is essential to ethics. It is a concept that—in its own way—reflects the fact that ethics are not guided by the inner self but by a sense of duty. In my view, it is absolutely central to morality [5], so allow me to say a little bit more about it here.

Counterfactuality is a very straightforward concept, and it underlies every little thing that we do every day: crossing the road without getting knocked over means counterfactually envisaging a situation where we are safe and sound on the other side. It is counterfactual because we have not crossed yet. To achieve our goal, we let ourselves be guided by a counterfactual ideal in which we are safe and sound on the other side of the road.

This example, which has no ethical implications at all, shows that even the act of crossing the road *is not simply an adaptation or reaction to external stimuli*. If we were only reacting to stimuli, we would not know what to do or where to go; we are guided not by our reactions but by a *counterfactual goal,* in this case that of arriving safely on the other side of the road. It is this element of counterfactuality that gives meaning both to our action and to the stimuli to which we must respond.

In this simple example, the element of counterfactuality is clearly determined: getting to the other side of the road. I have absolutely no doubt that a task such as this, which though complex has an unequivocal purpose, could be performed by a machine in the form of an algorithm since, in this case, the model is clearly the technical model of the task to be performed, that of finding an adequate means of going from A to B. In ethical situations, however, not only do we not know the *outcome,* given that we are in fact looking for the right thing to do, but also and above all—and this is decisive!—*the outcome will vary depending on the moral grammar we choose.* The end itself depends on the means chosen, and this is inherent in all moral reasoning. For example, in the 'simple' case of abortion (simple in the sense that the answer is either yes or no), the outcome will be completely different depending on whether we analyse the situation in terms of rights, values, utility or something else, not to mention the fact that these different moral grammars must be constructed from the perspective of each person involved (the mother, the child, the doctor, family and friends, etc.).

Such is the *indeterminacy in principle* of the entire space of reasons, on the grounds of which moral agents are likely to be able to justify and direct their actions. These 'grounds' are the huge counterfactual space from which agents draw their reasons. And they draw freely from them, so freely that, for human agents, acting morally implies that they can choose *not* to act morally. Yet this is precisely what we want to avoid with machines! The main aim of *machine ethics* is to program machines so that they *cannot* act immorally. While this is perfectly understandable, it means that we basically strip them of what gives us our autonomy, in the philosophical sense of the term. We just want them to follow a rule and, because we want to prevent them from deviating from that rule, we restrict their freedom of action.

So, by programming a moral grammar into the machine, a grammar that has been chosen by the programmer, we cut the machine off from the counterfactual space that gives human ethics its very substance. The machine will only ever be able to act according to the principles imposed factually on it by its program, whereas ethical human conduct draws necessarily on counterfactual resources—without which it is said that human beings would act 'like robots'…

As I understand it, there is a useful parallel to be made between the problem of robot locomotion, as described by Jean-Paul Laumond, and the problem of ethics. I have heard Laumond say that "the locomotion function was mastered perfectly from a reactive perspective, but not from a strictly decisional perspective, whereby it would be possible to string together several modes of locomotion, as we humans do every day" [6]. In other words, unlike a three-year-old child who is capable, in a new locomotion situation, of quickly leaning on an elbow for balance, grabbing a handrail, crawling, pushing obstacles aside with their foot or chin while at the same time spontaneously changing strategy, a robot is essentially reactive and therefore cannot yet function in an environment that requires a decision on the type of locomotion to employ. This spontaneous choice of locomotion mode corresponds analogically to the choice of moral grammars that we have at our disposal to address a problem. But there are two types of instructive differences with locomotion:

- First, with regard to locomotion, the objective is clearly identified, and the adequacy of locomotion modes is determined in relation to this objective. In ethics, as I have explained, the opposite is true: it is the means (that is, the moral grammars chosen) that determine the objective, in other words the right action to take; a utilitarian approach to a situation will not produce the same outcome as a deontological approach, etc.
- Second, with regard to movement, motor skills must be coupled with the sensory skills needed to estimate one's own position and angle, and hence make the appropriate movement. This sensory information consists of factual data that are collected and used in the context of the machine's motor abilities. However, we cannot simply rely on factual data to make moral judgements, as an item of data is relevant only in light of the counterfactual principle that explains it.

Therefore, the main problem in ethics lies in gaining access to counterfactuality. Until we have resolved this problem, we will have obedient machines but not ethical machines. In other words, the broad challenge facing *machine ethics* lies in accessing counterfactuality, or the 'grounds' on which an action should be carried out. These 'grounds' determine which information is relevant and why, and steer the action to be taken. We *choose* this counterfactual element, and it is this choice that determines the ethical quality of our action. Ethics is strictly inseparable from counterfactuality, whereas robots are riveted to the factuality of the algorithms implanted in them.

5 Conclusion

Let me sum up. In stage one, I demonstrated that the fact that programmers must choose a single ethical approach is reductive and does not reflect our ethical experience. In stage two, I explained the difficulties of principle associated with algorithmic formalisation, in terms of the action itself, its justification and its informational basis. Lastly, in stage three, I showed how ethics is inseparable from *counterfactuality*, which alone gives ethical meaning to the situations we find ourselves in. On this basis, I conclude first that AMAs as we know them today cannot be described as either agents or moral, but rather as *executors of pre-programmed rules*. Second, if agents are to be feasible, programmers of artificial ethical machines must solve the problem of their access to what I call counterfactuality.

The more general conclusion that I would draw from all of this is that we cannot talk about *machine ethics* indiscriminately, making do with a few vague notions that we have all had since childhood about what is ethical or not. Moral reasoning and judgement must be subject to careful analysis, which will reveal the full extent of the problem with artificial moral agents. Before we can discuss *machine ethics* or artificial ethics, we must, if we are to avoid speculation and ideology, have a clear understanding of what ethics is, what type of rationality it implements, and what is the nature of ethics and ethical conduct in general. Today's attempts to develop machine ethics are most certainly helping ethicists and philosophers to understand

all this, perhaps differently than if these attempts had never been made. Nevertheless, we simply cannot talk about machine ethics in any meaningful way unless we know what ethics is, just as we cannot make locomotor robots if we do not know what movement is.

References

1. Moor, J.H.: The nature, importance, and difficulty of machine ethics. IEEE Intell. Syst. **21**(4), 18–21 (2006)
2. Moor, J.H.: Four kinds of ethical robots. Philosophy Now **72**, 12–14 (2009)
3. Wallach, W., Allen, C.: Moral Machines: Teaching Robots Right from Wrong. Oxford University Press, Oxford (2009)
4. Anderson, M., Anderson, S.: Machine ethics: creating an ethical intelligent agent. AI Mag. **28**(4), 15–26 (2007)
5. Hunyadi, M.: L'Homme en contexte. Cerf, Paris (2012)
6. Laumond, J.-P.: Interview; La méthode scientifique, France Culture radio, 14 June 2017