Some Technical Challenges in Designing an Artificial Moral Agent

Jarek Gryz(✉)

York University, Toronto, Canada
jarek@cs.yorku.ca

Abstract. Autonomous agents (robots) are no longer a subject of science fiction novels. Self-driving cars, for example, may be on our roads within a few years. These machines will necessarily interact with the humans and in these interactions must take into account moral outcome of their actions. Yet we are nowhere near designing a machine capable of autonomous moral reasoning. In some sense, this is understandable as commonsense reasoning turns out to be very hard to formalize.

In this paper, we identify several features of commonsense reasoning that are specific to the domain of morality. We show that its peculiarities, such as, moral conflicts or priorities among norms, give rise to serious challenges for any logical formalism representing moral reasoning. We then present a variation of default logic adapted from [5] and show how it addresses the problems we identified.

Keywords: Moral agents · Artificial life

1 Introduction

When we evaluate software tools and systems with respect to fairness, accountability, and transparency, we engage in their moral evaluation. We assume, at least for now, that such an evaluation is performed by a human, that is, the system itself is not capable of moral reasoning. However, with the rapid progress in the design and development of autonomous agents this assumption may no longer be true. Not only do we need to evaluate software systems from a moral perspective, these very systems will very soon need to perform moral reasoning themselves. One can imagine a self-driving car facing an inevitable collision in which either a pedestrian or an occupant of the car will be killed. Assume furthermore that the only thing the driving agent can do is to maneuver so that only one of them dies. This is a moral decision and it has to be made instantaneously by the agent itself.

The urgency of the design of moral agents has not been lost on AI community. In 2017 at PRIMA a full day was devoted just to the topic: "Can we, and should we, build ethically-aware agents?" As should be clear from the previous paragraph, we believe that this is no longer a question of *whether* but *how*. We need a formal mechanism that will allow an autonomous agent perform moral

L. Rutkowski et al. (Eds.): ICAISC 2020, LNAI 12416, pp. 481–491, 2020.
https://doi.org/10.1007/978-3-030-61534-5_43

reasoning and reach the same (or at least very similar) decisions that a human being would reach in the same circumstances.

There are generally two approaches to design such a mechanism: bottom-up or top-down [12]. In the bottom-up approach, we try to mimic child development: human beings do not enter the world as competent moral agents yet learn enough of that competency within a few years. There have been several projects to simulate artificial life or emergence of social values and one would hope that similar approach would work in developing/simulating morality. There are many unresolved issues with this approach, however. First, psychologists disagree how much influence nature vs. nurture play in developing a theory of morality in children. Second, they disagree what the guiding principle in this development is: reason or empathy. Third, children are subjected to reward and punishment (approval-disapproval) from the society when learning moral behavior; it is not clear what would correspond to those in a machine. In the top-down approach, ethical principles and rules are explicitly stated and an agent simply follows them via an algorithm. Although this approach has its own problems – which we discuss in Sect. 2 – we believe it is more likely to succeed than the bottom-up approach.

The paper is organized as follows. In Sect. 2, we point out several technical challenges in designing and implementing a formalism for moral reasoning in an autonomous agent. In Sect. 3, we present a variation of default logic extended to moral reasoning, which addresses most of the problems we identified. This formalism has been adapted from [5] where it was developed as a solution to practical problem of justifying one's actions, that is, providing reasons for these actions. We conclude with a discussion in Sect. 4 where we point out that default logic still falls short of adequately representing human reasoning as it fails to capture the holistic and open-ended character of such reasoning. We need an empirical study of an actual implementation of a moral agent to determine how serious this issue is. We leave it, however, as an open question in this paper.

2 Challenges

2.1 Choice of Ethics

The first decision a designer of an artificial moral agent faces is the choice of an appropriate theory of normative ethics. Such a theory provides foundations for moral reasoning of the agent. There have been numerous proposals for ethical theories in the history of moral philosophy but the majority of them fall into one of just three categories: consequentialism, deontological ethics, and virtue ethics. Consequentialism emphasizes the consequences of moral actions; deontology emphasizes duties or rules, and virtue ethics virtue or moral character of an agent. The vocabularies used in each of these theories are different too. The language of consequentialism talks of "benefits", "outcomes", "pleasures" and "pains" and they refer to the result of a moral act. The terms from the second set serve in the prescriptive function of a moral code. This function consists in providing answers to questions like: What am I (morally) required to do?

Answers to such questions usually have the grammatical form of an imperative and are called 'prescriptions', 'moral norms', 'rules', 'precepts', or 'commands'. They are expressed by means of such terms as 'right', 'obligation', 'duty', etc. The third class contains terms used for a moral evaluation of an action (or an actor). Terms used for evaluations include 'good', 'bad', 'blameworthy', 'praiseworthy', 'virtuous', etc. Consider how an obligation to keep promises is justified within each of these ethics. A consequentialist might say that keeping promises increases trust in society, which benefits everyone. A deontologist would point out to a duty – stemming from some higher-level rule (e.g. "Do unto others as you would be done by") or perhaps from some religious authority – that we as humans are obliged to observe. A virtue ethicist will emphasize the good character of a person and say that keeping promises is something that a virtuous person would do.

The three theories tend to agree – in most cases – when evaluating moral decisions as right or wrong; after all, they have to agree with our moral intuitions to be acceptable. Still, we should assert again that a designer of an autonomous moral agent faces a critical decision in choosing the theory. The transparency requirement stipulates that we explain why an agent chose a particular course of moral action. So even though all three theories might tell the agent to do the same thing, the reasons for doing so would be radically different. Whether or not these explanations are acceptable by the humans interacting with the agent may thus very well depend on the choice of a theory.

All three theories have been formalized using appropriate logics thus allowing moral reasoning for each ethics (most recently in [2–4]). Pros and cons for each of these three theories have been debated in philosophy for ages. Needless to say, we are not going to engage in this discussion here. In fact, to keep the focus in the paper, we are going to discuss only the deontic ethics as the one, which is the most appropriate for implementation for a moral agent. We have two brief arguments against the competing theories. First, virtue ethics seems out of place in a realm of non-humans: how can one talk about a "good character" or "virtue" of a robot? These terms seem to be strictly reserved for humans; we do not even ascribe them to animals (other than in metaphors). Second, although consequentialism seems the easiest to formalize, computing and ranking utilities or benefits of all possible actions seems hopeless (in fact, incommensurability of the outcomes of actions was one of the main arguments against consequentialism). Thus, in the rest of the paper we discuss the possible implementation of the deontologic ethics only. Still, our choice is somewhat arbitrary so the first challenge is: what ethical theory should an autonomous agent use?

2.2 Deontic Logic and Moral Dilemmas

Historically, the most popular way of formalizing deontological ethics has been via deontic logic. The formalism introduces two operators O and P, which represent obligation and permission respectively. Thus, a statement $O(A)$ means that it is obligatory that A (or it ought to be the case that A). One can also talk

about conditional obligations: $O(A/B)$ means that under the circumstances B, it ought to be the case that A.

Although deontic logic has been widely used to formalize moral reasoning, it has a surprising weakness: it rules out the possibility of moral dilemmas. Yet it seems that we often face such dilemmas in our life. Sartre [9] tells of a student whose brother had been killed in the German offensive of 1940. The student wanted to avenge his brother and to fight forces that he regarded as evil. However, the student's mother was living with him, and he was her one consolation in life. The student believed that he had conflicting obligations. Yet when we formalize these obligations in deontic logic as $O(A)$ and $O(\neg A)$, we derive - by standard semantics of that very logic - a statement $O(A \wedge \neg A)$ which is unsatisfiable.

There is still disagreement among moral philosophers whether moral dilemmas might arise. Some of them would claim (we discuss this position in the next section) that what we see as conflicting obligations can always be prioritized; after all, we always manage to choose one over another. Still, it seems an awkward decision to build a philosophical position on moral dilemmas into the logic itself.

Thus, the second challenge is: if we believe that moral dilemmas are real and unavoidable, we need a formalism that represents them (deontic logic cannot).

2.3 Priorities Among Obligations

One can take a strong philosophical position and deny the validity (or at least likelihood) of moral dilemmas. Consider two such norms: "Keep promises" and "Save human lives" and imagine you are walking to teach a class and you see a drowning child in a nearby river. On one hand, you made a promise to your students and university administration that you will teach the class, on the other hand you should try saving the child. You cannot do both. Nevertheless, one can argue that this example does not really represent a moral conflict. Indeed, most likely everyone would agree that saving a child trumps keeping a promise (of any kind). Thus, we can introduce preference order between prima facie oughts and mark it as $A \leq B$ with the meaning that B is more important than A. Does the preference order have a property of strong connectivity, that is, either $X \leq Y$ or $X \geq Y$ for any arbitrary obligations X and Y? This strong connectivity would allow a convenient resolution of any moral conflict. Unfortunately, there is no agreement among moral philosophers whether all obligations are comparable. Some philosophers claim that any moral conflict can be resolved while others think that moral dilemmas are real and unavoidable. Most of us faced moral choices where it was not clear at all which obligation was stronger so our ordinary intuitions tell as that at least some obligations are incomparable.

However, even if we believe that all obligations are comparable, we can still face a moral conflict of two obligations that cannot be satisfied at the same time. Think of a single norm that is a source of two obligations. In the example of the self-driving car, there is one such norm, "Save human life" that is a source of two distinct obligations "Save person X" and "Save person Y". How do we handle

this problem for an autonomous agent? The agent can either make no choice between the two obligations (do nothing) or choose randomly between them (toss a coin). Neither of these options seems satisfactory. The first option may simply be not available to an agent. In our example the self-driving car has to choose who dies - it does not have an option of not deciding. On the other hand, if the agent chooses randomly, its moral reasoning is, legally speaking, deplorable and unacceptable. We do not leave moral decisions to chance. It seems then, that we must rank all obligations in a strict order so that an agent always has a clear choice of action.

An interesting aspect of our moral life is that we are able to offer reasons why we give preference to one obligation over another. And often, due to these higher-level reasons, we reverse previously held preferences. For most of us, the obligation "You shall not kill" is likely to have the highest priority among all moral obligations. However, even that one – again, for most of us – loses its status at the time of war or in self-defense. So perhaps, we should allow the strict preference among obligations to be dynamic as well, that is, allow some flexibility with respect to the context in which an agent operates.

Thus, the third challenge is: how do we (or can we) arrange all obligations an agent may encounter in a strict preference order that can be dynamically adjusted in different contexts?

2.4 Moral Reasoning Is Defeasible

One of the most interesting features of moral reasoning is its defeasible character. Consider again the example when I walk through campus to teach a class. I have not reached the river yet and the only obligation I am under now is to keep the promise to start my lecture on time. When I reach the river, I notice that a child is drowning. I redo the moral reasoning, reach a new conclusion that my obligation is to save the child and withdraw the conclusion about the obligation to keep the promise to teach the class. In other words, adding a new premise to my reasoning (a proof) made me abandon the previous conclusion. We use this type of reasoning almost every day; we may hold a certain moral opinion about some event only to change it in light of new facts.

There is yet another way our reasoning can be defeasible. The general rule describing the obligation to save human lives clearly has some exception (this is precisely why this is a general rule and not a hard obligation). When I walk by a drowning child and I see the police already at the scene I am no longer obliged to assist. I should also not save the child when I will put my own life in danger or when I cannot physically get to the river due to a physical barrier, etc. Every moral norm has exceptions. We are obliged to follow a norm unless and until we learn that an exception to the norm applies.

This is rather unusual for logic. The consequence relation of a classical logic is monotonic: if a formula p is a consequence of a set of formulas S, then p is also a consequence of $S \cup \{r\}$, for an arbitrary formula r. In other words, the set of conclusions we can draw from the premises grows monotonically with an addition of new formulas to the premises. In particular, a proof of a formula

cannot be invalidated by adding a new formula to the derivation. But this is not the case in common sense reasoning, in particular, moral reasoning. This type of reasoning cannot be captured by classical or even modal logic, which is often taken as a foundation of deontic logic. We need a different type of logic where monotonicity no longer holds.

Of course, we can avoid this challenge when the world of the autonomous agent is completely static, that is, no updates take place in the database describing the world around the agent. Clearly, this is not a realistic assumption for most applications.

Thus, the fourth challenge is: how do we formalize defeasibility of moral reasoning?

2.5 Ought Implies Can

One cannot expect anyone to do something impossible. Kant formulated this principle in the context of ethics as: "The action to which the "ought" applies must indeed be possible under natural conditions". In other words, if I am obliged to save a drowning child, I must be able to do so in a particular situation. It is surprisingly difficult, however, to specify what conditions have to be satisfied to make me able to save the child. For example, I cannot be handicapped, I have to have access to the river, I have to be able to swim well, the river should be slow moving rather than a torrent, etc. It is pretty much impossible to tell when this list is complete. Indeed, this problem has been identified many years ago in the context of planning in AI as the qualification problem [6]: to plan an action we need to know what initial conditions have to be satisfied for this action to succeed. The famous example in that context was the problem of necessary conditions to start a car: the battery is charged, there is gas in the tank, the exhaust pipe is not blocked, nobody has stolen the engine at night, etc. Ordinary logic could not solve that problem for actions in general and it is unlikely that it could do so for moral actions in particular.

Thus, the fifth challenge is: what practical conditions have to be satisfied to say that an agent is under a moral obligation to perform a certain action?

3 Default Logic to the Rescue

3.1 Default Logic

Default logic was originally proposed in [7] to solve planning problems in classic AI. The idea behind default logic was to account for some aspects of our commonsense reasoning. We tend to learn about the relationships in the world by making sweeping generalizations, such as all swans are white or all birds fly. And then, when we see a black swan or learn about ostriches, we retract or qualify our previous claims. In this sense, common sense reasoning is non-monotonic: a conclusion set need not grow monotonically with the premise set. If we could formalize this type of common sense reasoning, then we might be able to account for intricacies of moral reasoning.

In addition to standard rules of inference, default logic adds default rules, which represent defeasible generalizations. A default rule has the form $\alpha \rightarrow \beta$, where α is a premise and β is the conclusion. The meaning of the rule is: if α has been already established, one can add β to the set of conclusions assuming that this set is consistent with β. A default theory is a pair $\Delta = <W, D>$, in which W is a set of ordinary formulas and D is a set of default rules. Consider again the generalization "all birds fly". This can be represented as a default rule, such that $\sigma = B(x) \rightarrow F(x)$ with the meaning if x is a bird, then x flies unless we have information to the contrary. Thus, if all we know about *Tweety* that it is a bird, we conclude that *Tweety* flies. Once we learn, however, that *Tweety* is an ostrich (hence does not fly, formally, $\neg F(Tweety)$), we cannot draw a conclusion that *Tweety* flies, as it is inconsistent with what we already know. Within AI, default rules were designed to address the qualification problem, the problem of formulating useful rules for commonsense reasoning amidst a sea of qualifications and exceptional circumstances. Going back to the example from Sect. 2.5, we want to be able to say that turning the key starts the car without having to specify all the exceptions to the rule. If any of these exceptions do occur, they will simply block the application of the default (just like *Tweety's* being an ostrich blocks the rule that it flies).

To accommodate new rules of inference, the standard concept of logical consequence has to be modified.

Definition 1. *The conclusion set Γ associated with a default theory $\Delta = <W, D>$ is called an extension and is defined as a fixed point:*

$$\Gamma = \sum_{n=1}^{\infty} \Gamma_i$$

where:

$$\Gamma_0 = W$$
$$\Gamma_i = Th(\Gamma_i) \cup \{B | A \rightarrow B \in D, A \in Th(\Gamma_i), \neg B \notin \Gamma\}$$

and Th(i) is a set of standard logical consequences of i.

The idea behind this definition is that we first conjecture a candidate extension for a theory, Γ, and then using this candidate define a sequence of approximations to some conclusion set. If this approximating sequence has Γ as its limit, Γ is indeed an extension of the default theory.

A default theory can have multiple sets of conclusions, that is, extensions. A famous example [8] of this case is called the Nixon Diamond: Nixon is a republican but he is also a Quaker. Republicans tend not to be pacifists and Quakers tend to be pacifists. As a default theory, these facts can be stated as: $W_1 = \{Q(Nixon), R(Nixon)\}$ and $D_3 = \{\sigma_1 : Q(x) \rightarrow P(x), \sigma_2 : R(x) \rightarrow \neg P(x)\}$. This theory has two extensions, one with $P(Nixon)$ and one with $\neg P(Nixon)$. Both conclusions are equally valid, yet they cannot be both entertained at the same time. This seems like a natural description of commonsense reasoning which cannot be captured in classical logic.

This formalism can be enriched by adding priorities between defaults. If we believe that being a republican is an extremely strong indication of being non-pacifist (at least much stronger than being a Quaker is an indication of being a pacifist), then we can state that $\sigma_1 < \sigma_2$ with the meaning that if these two default rules apply at the same time, only the second will fire. Our extension will contain only the fact that Nixon is not a pacifist.

3.2 Obligations in Default Logic

How exactly do we represent formally obligations in default logic? Let $O(A)$ be an obligation "You should do A".[1] This can be represented as a default rule $\sigma : \mathbf{T} \rightarrow A$ ($\mathbf{T}$ stands for tautology which means that the obligation is unconditional).

Definition 2. *Let $\Delta = <W, D>$ be a default theory and the default, $\sigma : \mathbf{T} \rightarrow A \in D$ represents an obligation. Then $O(A)$ follows from Δ just in case $A \in \Gamma$, for some extension Γ of this theory.*

Default logic can represent conflicts between obligations in a straightforward way. Consider again the example of encountering a drowning child on the way to a lecture. We can represent two relevant obligations as default rules $\sigma_{life} = O(L)$ and $\sigma_{promise} = O(P)$ with the meaning respectively "you should save human lives" and "you should keep promises". We are assuming here, of course, that these two obligations are logically incompatible in this particular context, which can be expressed as $O(P) \Rightarrow O(\neg L)$ and $O(L) \Rightarrow O(\neg P)$.[2] If we do not prioritize between these two default rules our theory will have two incompatible extensions (just like in the Nixon example), one telling us to save the child, $O(L)$, the other one telling us to walk to the lecture, $O(P)$. On the other hand, if we prioritize between these two default rules by saying that saving human lives is more important than keeping promises, that is, $\sigma_{promise} < \sigma_{life}$, then we will only have one extension containing the statement that you should save the child.

We assumed so far that the priority relations among default rules representing obligations are fixed in advance. As discussed in Sect. 2.3, however, we would like to have some flexibility in setting these priorities, that is, to be able to adjust the order depending on a context. Default logic offers a straightforward mechanism to do just that. Instead of stating $\sigma_1 < \sigma_2$ as a matter of fact, we can add it as a default rule to the set of other defaults. Formally, we would express it as $\mathbf{T} \rightarrow \sigma_1 < \sigma_2$ with the meaning "Obligation σ_2 has a higher priority than obligation σ_1 unless we have the information to the contrary" .

Default logic allows us to represent defeasibility of moral reasoning in two different ways. First, an obligation may be blocked by exceptions. I am under an obligation to save a drowning child unless I know that an exception to that rule applies, for example, the police are already at the scene. The concept of default extension from Definition 1 ("the rule applies unless I have information to the

[1] We rely on [5] for ideas and formalism of this section.

[2] We use '$\Rightarrow$' for standard logical implication.

contrary") conveys exactly that intuition. The second type of defeasibility arises through a dynamic change of applicable obligations (default rules). Consider again the scenario described in Sect. 2.4. Initially, $\sigma_{promise}$ is the only default rule that applies (we have not yet seen the drowning child) and we keep walking to the lecture. Then, when our database of facts gets updated and σ_{life}, which has a higher priority than $\sigma_{promise}$ applies as well, we retract the obligation that we should walk to the lecture from and introduce instead the obligation of saving the child.

To summarize, under the interpretation of obligations in default logic, we can account for:

1. Moral conflicts (challenge 2): incompatible obligations lead to multiple incompatible extensions.
2. Priorities among obligations (challenge 3): they are represented as priorities among default rules.
3. Defeasibility of moral reasoning (challenge 4): exceptions will prevent a default rule from firing a default rule with a higher priority will invalidate (make it inapplicable) another default rule with a lower priority.
4. Qualification problem (challenge 5): default logic was originally introduced to handle this problem (we assume an action is doable unless we have information to the contrary).

4 Discussion and Open Issues

Default logic is not a perfect solution to codifying commonsense reasoning. It solves many problems that classical logic could not, yet it leads to some counterintuitive results in other cases. Consider the following example that illustrates the famous *multiple extension problem*. Tom is a spy. As a human being, he ought to tell the truth. However, he is also a spy and spies routinely lie (or at least are not expected or required to tell the truth). Common sense would tell us that Tom is not required to tell the truth. The fact that he is a spy is more relevant to what we expect him to do than the fact that he is human. According to default logic, however, both extensions (one with Tom required to tell the truth and the one without this obligation) are equally valid. What was considered an advantage in the case of Nixon Diamond (both extensions seemed reasonable) is clearly a flaw here. Default logic in the form described above does not distinguish between extensions unless we put priorities between defaults. Thus, consider prioritizing between these two defaults by making lying more important to spies than telling the truth to humans. But is it always more important? Should Tom lie to his physician about his health or to his wife about picking up their son from school? Clearly, priorities between defaults/obligations do not hold universally but *depend* on context. We need to be able to tell when Tom's being a spy – hence his non-obligation to tell the truth - is relevant for each context. But the concepts of context or relevance cannot be codified in logic and implemented in a software system.

Let us take stock. We proposed default logic as a way of implementing moral reasoning. This formalism accounts for a number of features typical of moral reasoning that other types of logic, such as deontic logic, cannot handle. Yet we also discovered that the conclusions reached via default logic are sometimes counterintuitive and only an appeal to a context or relevance can provide an intuitive and correct result. So how do we, the humans, discover the relevant context? Most of it we probably learn from experience, some of it may be innate. The crucial question for AI is how this knowledge is stored and processed. In great majority of our actions, we do not consciously think through our plans. In fact, the relevant knowledge rises to the surface only when we make mistakes and need to reconsider our course of action or plan an entirely novel action. But even in these cases, introspection is not of much use. We do not ponder upon facts one by one; we are somehow capable to pay attention only to the relevant ones. Also, our information processing is so fast that it cannot involve drawing thousands of interim conclusions. We operate on a portion of our knowledge at any moment and that portion cannot be chosen by exhaustive consideration. A great discovery of AI is the observation that a robot (a computer) is the fabled tabula rasa [1]. For it to operate in a real world, it must have all the information explicitly specified and then organized in such a way that only the relevant portion of it is used when it is time to act. So far, we have no idea how to do that. The problem, in a nutshell, is this: "How is it possible for holistic, open-ended, context-sensitive relevance to be captured by a set of propositional, language like representation of the sort used in classical AI?" [10].

What then are the chances of building a moral machine? We believe that the conclusions should not be all negative. AI faces the problem of the holistic and open-ended reasoning only for the holistic and open-ended environment. However, when it restricts its attention to microworlds, that is, well defined and fully described cuts or aspects of the world it works very well. In fact, recent successes in areas such as face or voice recognition systems are already a source of much anxiety because they work so well. Our conjecture then is this: if we restrict the domain of an autonomous agent to a well-defined environment, it has a good chance of working correctly in that environment. In particular, if we specify the relevant knowledge necessary for moral reasoning in that environment one may expect to circumvent the multiple extensions problem. Nonetheless, only an empirical evaluation of an actual system implementing default logic can tell how close we are to building a moral machine.

References

1. Dennett, D.: Cognitive wheels: the frame problem in AI. In: Pylyshyn, Z. (ed.) The Robot's Dillema, pp. 41–64. Ablex Publishing Corporation (1987)
2. Oesterheld, C.: Formalizing preference utilitarianism in physical world models. Synthese **193**(9), 2747–2759 (2015). https://doi.org/10.1007/s11229-015-0883-1
3. Gabbay, D., Horty, J., Parent, X., et al.: Handbook of Deontic Logic and Normative System. Blackwell, Oxford (2013)

4. Govindarajulu, N., Bringsjord, S., Ghosh, R., Sarathy, R.: Toward the engineering of virtuous machines. In: AIES Conference (2019)
5. Horty, J.: Reasons as Defaults, Oxford (2012)
6. McCarthy, J.: Applications of circumscription to formalizing commonsense knowledge. Artif. Intell. **28**, 86–116 (1986)
7. Reiter, R.: A logic for default reasoning. Artif. Intell. **13**, 81–132 (1980)
8. Reiter, R., Criscuolo, G.: On interacting defaults. In: Proceedings of IJCAI (1981)
9. Sartre, P.: Existentialism is a Humanism. Meridian, New York (1957)
10. Shanahan, M.: The frame problem. plato.stanford.edu/entries/frame-problem. Accessed 7 Mar 2020
11. Touretzky, D.S.: Implicit ordering of defaults in inheritance systems. In: AAAI Conference (1984)
12. Wallach, W., Allen, C.: Moral Machines. Oxford University Press, Oxford (2009)