Automating Moral Reasoning

Marija Slavkovik
University of Bergen, Norway

Abstract

Artificial Intelligence ethics is concerned with ensuring a nonnegative ethical impact of researching, developing, deploying and using AI systems. One way to accomplish that is to enable those AI systems to make moral decisions in ethically sensitive situations, i.e., automate moral reasoning. Machine ethics is an interdisciplinary research area that is concerned with the problem of automating moral reasoning. This tutorial presents the problem of making moral decisions and gives a general overview of how a computational agent can be constructed to make moral decisions. The tutorial is aimed for students in artificial intelligence who are interested in acquiring a starting understanding of the basic concepts and a gateway to the literature in machine ethics.

2012 ACM Subject Classification Computing methodologies → Philosophical/theoretical foundations of artificial intelligence

Keywords and phrases Machine ethics, artificial morality, artificial moral agents

Digital Object Identifier 10.4230/OASIcs.AIB.2022.6

Category Invited Paper

1 Introduction

Artificial intelligence (AI) is concerned with the problem of using computation to automate tasks that require intelligence [13]. Artificial intelligence, since 1956 when it was named and established [35], has been increasingly contributed towards automating tasks that require manipulation of information, production line tasks (robotics) and most recently pattern identification and learning [15].

In a society, we all affect each other with our activities and decisions. Ethics (or moral philosophy) is concerned with understanding and recommending right and wrong behaviours and decisions [25]. The right decisions being characterised by taking into consideration not only ones own interest, but also the interest of others [29]. The more computationally automated tasks are used to complement or replace people's tasks, the more concerns we have to ensure that the resulting actions and choices are not only correct and rational, but also do not have a negative ethical impact on society. As Rosalind Picard puts it "The greater the freedom of a machine, the more it will need moral standards" [42]. AI Ethics is a new, interdisciplinary, sub-field of AI that aims to address precisely this issue.

One way to ensure that AI has a non-negative ethical impact on society is to ensure that we do have an insight into, and measures to control the impact AI has [22]. Various different research approaches are being developed towards this end, in computer science, but also in philosophy, organisational science, law etc. Algorithmic accountability studies how to ensure that society and stakeholders can establish the right relationship with the people who research, develop, deploy and use AI algorithms [53]. Transparency is concerned with ensuring that the adequate type of information about how an AI system works is made available to a given stakeholder [20, 54]. Fairness is concerned with ensuring that like individuals and groups are treated alike by decision-making algorithms [17]. Explainable AI is concerned with the problem of finding ways to extract information from AI algorithms that justifies the choices that algorithm takes and use that information to adequately explain that information to a given stakeholder [34, 28].

Another way to ensure AI has a non-negative ethical impact is to consider that moral reasoning is itself a cognitive task that we can consider automating. Machine ethics, or artificial morality, is a sub-field in AI that is researching this approach. In general, machine ethics is "is concerned with the behaviour of machines towards human users and other machines" [5]. The problem of automating moral reasoning can be considered as a problem of moral philosophy, whereas one is interested in questions such as: should machines be enabled with ethical reasoning [24, 8], which norms should machines follow [33], can machines ever be moral agents [16], etc. As a problem of computer science, machine ethics focuses on the question of how to automate moral reasoning [6, 51].

Here we are concerned with the question of how to automate moral reasoning. Although this problem, and machine ethics in general, have been raised since 2006 [5], it is an extremely difficult problem that requires a lot of improvement in the state of the art in AI and moral philosophy. We discuss the basic approaches in machine ethics, the advantages and challenges of each. These lecture notes are structured as follows.

We start with Section 2 in which we discuss what is decision making and how decision-making is distinguished from moral decision-making. Decisions are made by an agent. In Section 3 we discuss what computational agents are, what does it mean for a computational agent to be autonomous and what kind of moral agents can computational agents be. One way to automate moral reasoning is to follow a specific moral theory. In Section 4 we give a very quick overview of what is a moral theory and some of the more known moral theories from moral philosophy. In Section 5 we discuss two general approaches to building artificial moral agents, we discuss open research problems and challenges. In Section 6, we end with a discussion on how to find out more about machine ethics, beyond the scope of this tutorial.

2 Moral decision making

Decision-making is a cognitive process that is studied from many disciplines including cognitive science, neuroscience, psychology, and economy. Decision theory is a field which studies the choices that an agent does (descriptive decision theory) and should (prescriptive decision theory) make when faced with a formally specified decision problem [41]. Artificial Intelligence typically follows the model of decision-making from economy since the goal typically is to automate rational decision-making [45]. We summarise that model.

As [31] put forward, decision-making is taken to comprise of at least four steps:

1. identify the problem for which a decision needs to be made,
2. evaluate the objectives and preferences that apply,
3. analyze the decision problem and its constraints, and develop or identify the possible options from which to choose,
4. choose from the identified options following some reasoning.

To put this into perspective, consider an example. In step one, I as an agent recognise that I am hungry and that this is a problem. Next, I evaluate my objectives and preferences. So my objective here are to stop being hungry which means get something to eat. My preference would be to eat a hot meal but not cook and at the same time, to eat something soon. Thus my problem becomes finding a place to buy food from. I now (step 3) have to identify the constraints: how much I want to spend, how far I am willing to go and which establishments are open today. These constraints help me make a list of possible restaurants to choose from. In the last step I choose from among the alternatives using some type of reasoning, like for example the closest and cheapest Indian food restaurant.

Decisions are taken by an agent. In computer science artificial intelligence[1], the requirement of what constitutes an agent is very low. Anything that has the ability to perceive its environment through sensors and can act upon that environment through actuators is considered to be an agent [45]. For a computational agent the environment is essentially a data construct consisting of lists of information with "sensors" being various dynamic inputs that change, add or remove data available to the agent. The source of that input can be an actual device that measures aspects of the environment and converts it into data or it can be a human inputting the data directly. An actuator is a device that produces motion and changes a physical environment in that way. For a computational agent, the "actuators" are its ability to change its own environment, namely its ability to alter data that is available to not only themselves but also some other person or agent.

The difference between a computational agent and an embodied agent (such as a robot) is perhaps best illustrated by considering as an example a program that plays chess. While the chess playing program does not "see" a chess board, or physically move chess pieces, it still plays chess which is considered an activity that requires intelligence. The environment it works with is a digital representation of the chess board and the changes on it. What is important in playing a move is not that the chess piece is physically moved on the board, but to make the choice of which chess piece to move in order to win.

The problem identification, the evaluation of objectives, preferences and constraints that apply is part of situational awareness and for most computational agents this is supplied as data. Situational awareness is the perception of one's environment, the events in that environment with respect to time or space, internalisation and utilization of that perception as information, and the projection of the future state of the environment and its elements. The available options and their characteristics are often also made available with the agent expected to do the evaluation of the options towards finding the optimal choice with respect to the given objectives, preferences and constraints. While people are capable of reiterating steps 3 and 4 of the decision-making process by identifying missing information and procuring it, those same activities, as well as situational awareness in general, are a hard problem in AI [45].

What is the difference between decision-making and moral decision-making? The difference is in whose objectives, preferences and constraints we choose to apply. As we can see from the four step model, decision-making is a process that only considers the objectives, preferences and constraints of the agent that makes the decision. Moral decision making requires us to consider the objectives, preferences and constraints of others. For example, when making a list of options, the environmental impact of different food options can be considered. Although I strongly prefer meat to tofu, I can choose to rank the vegan options over the meat options because sourcing that food does not cause suffering to animals.

It is, of course, in general not clear how that information on the objectives, preferences and constraints of others is to be sourced, to which extent it should be considered etc. In moral philosophy, numerous different approaches to answering these questions for human decision-makers have been discussed and we will give a brief overview of some of the dominant ones in Section 4. In the next section, we will consider artificial moral agency: to which extent *can* a computational agent make moral decisions given that they do not have full agency to do so due to the lack of situational awareness and other constraints on information processing capabilities.

[1] Artificial Intelligence can be studying in other fields than computer science such as philosophy

Before we move on, although it is not directly relevant here, for completeness, it is important to mention the relationship between economy and moral philosophy. Rational decision-making is studied in economy, whereas what are good and bad decisions is studied in moral philosophy. One dominant perspective on moral decision making in economy is that of Sen [46]. In economy, a decision problem is represented with a set of available alternatives, and the agent's preference order over those alternatives. The decision-making process is then choosing the alternative that maximizes the expected utility for the agent. Sen [46, 47] attempts to model morality by assuming morality to be an ordering over the agent's preference of alternatives, namely an ordering over orderings. So a moral decision is then to choose which preference order over the available alternatives to follow.

Dietrich and List [21] also present a decision-theoretic consideration of moral decision making. They show how to represent a moral theory in terms of two parameters: "(i) a specification of which properties of the objects of moral choice matter in any given context, and (ii) a specification of how these properties matter." We discuss what moral theories are in Section 4.

3 Artificial Moral Agents

Autonomy is the ability of an agent to govern itself, which includes its ability to identify problems and make decisions to resolve them. Both [36] and [51], some of the pioneers of machine ethics, observe that the extent to which an artificial agent would be able to do moral decision-making depends on the extent of autonomy of the agent. We typically do not talk about the autonomy of artificial agents, but the level of autonomy of a system, which we then refer as an autonomous system. Before introducing the different types of artificial moral agents that [36] and [51] have proposed, we introduce the levels of autonomy of an autonomous system.

An autonomous system is a system, software or device alike, that is capable of some degree of operation without human control. Systems that do not have autonomy are divided into: controlled, supervised and automatic. Controlled systems are systems that have no autonomy and require continuous human control to operate. An example of a controlled system is the standard electrical iron for clothes: for the iron to be used, a person must hold plug it in, hold it and move it. Supervised systems are systems that are capable of some short periods of unsupervised activity, but require a human operator to both start and end that period of activity. An example of a supervised system is the standard washing machine. A person needs to load it and choose a program, and it operates without oversight executing the washing program, and the human operator needs to unloaded after the program is completed. Automatic systems are able to operate without any human supervision, but they can execute only a very limited range of activities in a fully controlled environment: all their choices are pre-programmed. An example of an automatic system is an elevator. It responds to a human input in a specified way, but it operates (moves up, down, stops, opens and closes doors) without the oversight of human operator.

Autonomous systems are systems that are able to operate without human oversight for long periods of time, they are able to process signals from the environment, use them to reason about their choices and actions, and be able to perform actions that have an effect on the environment. In autonomous systems we discern, according to Parasuraman et al. [40], ten levels of autonomy of decision and action selection (ranked here from high to low), listed as follows.

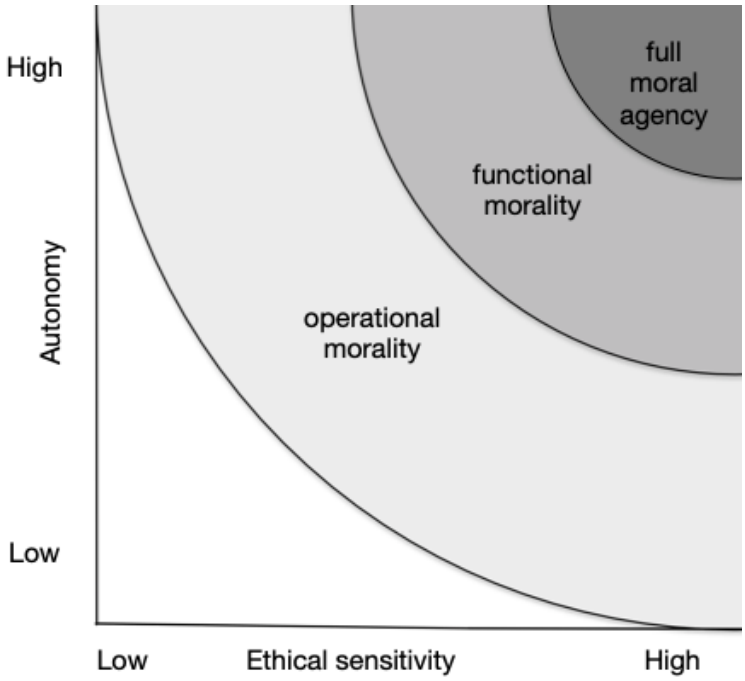

Figure 1 Degrees of artificial morality [51, Chapter2].

1. The system decides everything, acts autonomously and can ignore human input and control
2. Informs the human about its choices only if it, the system chooses to do so
3. Informs the human about its choices only if asked
4. Executes its decisions autonomously/automatically and then necessarily informs the human about the decision-making process
5. Allows the human a limited time to veto a decision made autonomously before execution, or
6. Executes the decision only if the human approves it, or
7. Suggest a decision (an alternative) to the human
8. Narrows the selection of options to choose from to a human, or
9. The system offers a complete set of decision/action alternatives, or
10. The system does not make any decision-making, the human must make all the decisions and actions.

Wallach and Allen [51] observe that the ability of computational agents to make moral decisions is restricted by their autonomy and by their *ethical sensitivity*[2]. They offer a graph of different types of artificial moral agency, which we reproduce in Figure 1.

[51] define operational morality as the morality of the artificial agents for which "the moral significance of their actions lies entirely in the humans involved in their design and use". This means that the artificial agent itself does not make moral decisions, but (in matters of morality) follows the instructions of a human operator. Functional morality is defined

[2] [51] do not explicitly define ethical sensitivity, but it is understood that it refers to the ability of the artificial agent to take into account the objectives, preferences and constraints of other when making decisions.

as the property of artificial agents who have the ability to make moral decisions without direct instructions from humans. It is understood that this ability is contextual, namely only applies under given circumstances. Full moral agency is the ability to have situational awareness and fully autonomously make moral decisions.

[36] also offers a four tier distinction between artificial moral agents, but does not base it explicitly on levels of autonomy or "ethical sensitivity" of the agent. Instead, he considers choice making artificial agents whose operation affects others in society in a positive or a negative way. For these agents he considers whether the agent uses moral choice relevant information at all, and if it does whether it sources it itself or it is in some way provided. [36] discerns: ethical impact agents, implicit ethical agents, explicit ethical agents and fully ethical agents.

Ethical impact agents are artificial agents which by virtue of existing bring about positive or negative impact on the lives of people. An ethical impact agent is an autonomous system or an AI system that is a disruptive technology, namely it changes society, and makes life better or worse for people in it by changing how certain tasks or operations are executed. An ethical impact agent does not itself make moral decisions at all or considers the objectives, preferences or constraints of others in its decision-making.

Implicit ethical agents implicitly do moral decision making. Namely, they are either fully constrained from choosing unethical options or the options they are considering are made available already evaluated with respect to how they affect the objectives, preferences and constraints of others. The evaluation of what is right and what is wrong thus is performed entirely by the human designers or operators of the artificial agent.

Explicit ethical agents do explicit moral decision making. This means that they are able to evaluate if an option is more or less ethical than another, possibly by also sourcing their own information for this evaluation. The degree to which they can perform moral decision making, would of course depend on the limited abilities of the artificial agent and might also be contextual. The evaluation of what is right and what is wrong is still by a large extent determined by the information supplied by the human designers or operators, but the artificial agent also contributes.

Fully ethical agents for Moor [36], just like for Wallach and Allen [51], are agents who have situational awareness and fully autonomously make moral decisions.

The definitions of Moor [36] are not meant to be operational. It is in general very difficult to evaluate the impact of an artificial agent, and also to draw a line to indicate which agents have and which do not have such impact. A mobile phone can be considered an ethical impact agent: it helps to solve crimes (a net positive impact) and it eases surveillance of people (a net negative impact[3]). To be able to ascertain whether an agent is an implicit or explicit moral agent one necessarily needs to have access to the agent's programming. [23] propose to refine the taxonomy of [36] towards making it more operational.

[23] propose that implicit can be considered those agents who engage in moral decision making without using their own autonomy. For ethically sensitive contexts, implicit ethical agents defer to the human operator, either directly or by accessing information made available. Explicit ethical agents do use the autonomy they have to make moral decision. Both explicit and implicit ethical agents thus have sufficient situational awareness to recognise that moral decision making is needed. However, implicit agents, do not use their autonomy, even if it is high. Their moral choices are either constraint or otherwise governed by a human operator. In contrast, explicit moral agents do use the autonomy they have to make moral choices.

[3] Personal opinion

One idea on how to develop implicit or explicit ethical agents is having their moral choices be governed by theories developed by moral philosophy. The alternative is that the artificial agents are informed directly by a person or societies views on what is right and wrong, see for example [39, 11, 44].

4 Moral philosophy

A detailed overview of moral philosophy is outside of the scope of this lecture. Perhaps choosing which moral philosophy works to present and which to ignore in a short list, itself is a moral choice. Instead we give a definition of what a moral theory is and what the main types of moral theories have been considered in moral philosophy. Each of the theories have strengths and weaknesses. Furthermore, every theory is developed to be applied by a human agent.

Moral philosophy is considered to include three main areas of study: meta-ethics, normative ethics and applied ethics [25]. Meta-ethics is concerned with the concepts of right and wrong themselves and how the validity of these concepts can be established. Normative ethics is concerned with developing means to identify what are the right and wrong decisions, actions, states of the world etc. Applied ethics is concerned with issuing recommendations of what is the right thing to do for a specific person in a specific situation. An example of applied ethics are the biomedical ethics rules that govern the conduct of, among others, medical doctors [12].

In machine ethics, we are primarily interested in normative moral philosophy and the moral theories developed within it. A moral theory is an explanation of what makes an action right, or what makes an entity good [50]. It is a reasoning system that can be used to establish the righteousness of an action or the goodness of an entity etc. The moral theories that are concerned with the discerning between good and bad entities are called theories of value, whereas those concerned with discerning between good and bad choices or actions are called theories of obligation.

Many moral theories have been proposed. Vaughn [50] argues that for a theory of reasoning about right and wrong to be usable, it needs to satisfy the following basic criteria. It needs to be consistent with considered judgments and our moral experience. Considered judgements are morally relevant preferences or decisions society has already made by carefully considering the complexity of a given problem and the intended and unintended consequences of alternative options. Our own moral experience vaguely describes what most people have been raised to intuitively consider right or wrong in most situations, such as for example stealing or betraying a confidence. Further, a moral theory should be useful in moral problem solving and it should be coherent. Usefulness means that anyone can apply it to make moral decisions, whereas coherence means that it should identify the same moral choices when presented with the same problem features.

When a decision is made, three aspects of the decision can be considered to be most relevant for identifying if that decision is good or bad. These are: the agent that makes the decision (their intentions, objectives and incentives included), the decision itself and available alternatives, and lastly the consequences of that decision.

Moral theories that put most relevance on the properties of the agent that makes the decisions are called *virtue theories*. The theories which place most relevance on the decision and alternatives are called *deontological theories*. *Consequentialist theories* deem that what ultimately decides whether an action is good or bad is the consequences of that action.

Virtue theories prescribe not how to make a decision but what intentions, objectives and preferences, i.e. virtues, the agent should have in order to choose right. The moral conduct of the agent emerges from their moral virtues. A virtue is a stable disposition to act and feel according to some ideal model of excellence. A notable virtue theory is Aristotelian ethics. Aristotle claimed that intellectual virtues can be taught but that moral virtues can be learned only through practice. He argued that for an agent to be virtuous they have to aim to achieve the golden mean which is finding a balance between two behavioural extremes. While at first glance it may seem that virtue theories are not particularly suited for developing artificial ethical agents, this is not necessarily the case. For example, [48] argues how virtue ethics can be taken from theory to implementation.

Deontological theories prescribe that the righteousness of a decision should be based on whether the chosen option is itself right or wrong under a series of rules, rather than on who is executing it [2]. Deontological theories typically prescribe obligations, permissions, prohibitions that the agent should follow when choosing between alternatives. Deontological theories also prescribe ethical values or ethical principles that one should follow in order to identify the right choices. In a very abstract way, they can be seen as providing heuristics for what are the objectives, preferences and constraints of others that the agent should taken into consideration during moral decision-making.

A notable deontological theory is Kantian ethics. Kant argued that reason alone leads us to the right and the good. According to him, moral law is a set of imperatives one should follow: hypothetical or categorical. A hypothetical imperative tells us what we should do if we have certain duties (obligations), while a a categorical imperative tells us what we should do regardless of our wants and needs. For example, [43] and [14] argue how Kantian ethics can be used to develop artificial moral agents. It should also be noticed that deontic logic has been developed to formalise reasoning about obligations and norms [26].

Consequentialist theories are perhaps the first thing that does come to mind when one considered moral theories. These theories prescribe that a decision is moral if it is motivated by assessing the consequences of the available options, namely what kind of states of affairs they bring about [2].

The most notable consequentialist group of theories is utilitarian ethics. Utilitarianism is the theory asserting that the morally right action is the one that produces the most favourable balance of good over evil, everyone considered [50]. Act-utilitarianism is the theory that the morally right actions are those that directly produce the greatest overall good, everyone considered. Given that it is not always practical to assess the consequences of each available alternative, sometimes rules can be developed to identify the actions that typically have desirable consequences. Rule-utilitarianism is the theory that the morally right action is the one covered by a rule that if generally followed would produce the most favourable balance of good over evil, everyone considered [50]. Since utilitarianism essentially prescribes quantifying the goodness of an action and reduces moral decision-making to maximising the utility of an option, the idea that this moral theory can be implemented in an artificial agent is not strange and has been considered early on, see for example Gips [27] and Anderson and Andrson[4].

5 Top down and bottom up automation

How should one go about developing an artificial moral agent? The first choice is to asses to which degree can we predict the moral decision-making problems that the artificial agent is expected to handle and what is the impact of the choices that the agent will make on

the environment. This assessment will inform us towards whether we need to build an implicit or an explicit ethical agent. Beyond this choice, Wallach, Allen and Smit [52] argue that artificial moral agents can be built either by following a top-down or a bottom-up approach, not excluding hybrids of these two approaches as well. Top-down and bottom-up approaches are typical heuristics in problem solving deployed in engineering. Both top-down and bottom-up approaches can be executed for both implicit and explicit agents.

Following the top-down approach, a problem is iteratively broken down into smaller problems until we reach a problem small enough that we know how to solve. The solutions of those smaller problems combined constitute the solution of our original problem. Procedural programming operates following the top-down approach, where an algorithm breaks down the problem into a set of basic instructions that the computer can execute.

Following the bottom-up approach, we start by describing the features or criteria of the solution of the problem. Different actions are then pieced together, possibly in a trial-and-error fashion, until a solution is reached that satisfies the prescribed criteria. Declarative programming operates following a bottom-up approach where we describe the sought information or alternative by giving constraints and preferences that should apply to it.

Using the top-down to build an explicit ethical agent means answering the question: how can we build the artificial agent to follow a given moral theory? To build an implicit agent top-down means to answer the question: how can we follow a given moral theory to build an artificial agent? Both explicit and implicit agent construction requires facing the challenge of choosing a moral theory. Virtually all of the theories from moral philosophy can be difficult even for people to follow. For each of them there exist examples of situations, called moral dilemmas, in which the theory does not provide a satisfactory method to choose what to do. The approach then is not to attempt to fully implement a moral theory, or expect that the artificial agent will succeed in implementing it where humans have failed.

Building explicit top-down agents requires difficult AI problems to be solved, such as situational awareness, prediction of consequences of ones actions. So it is the state of the art in AI that is also a limitation to the abilities of artificial moral agents of this kind. Building implicit top-down agents is somewhat more attainable, but requires that the agent is given a pre-determine set of rules, built by a human operator following a moral theory, that they should apply when making moral decisions. This in turn, is only possible when the human operator can to a large degree predict the problems the artificial agents will face and the choices that would be available.

The advantage of the top-down approach is that the resulting agent, whether implicit or explicit, will follow a "tried and tested" theory. This makes it possible to test if the agent is making the correct choices according to some theory. Top-down approaches are also *verifiable* [19].

When building artificial moral agents, it is not sufficient to enable them to make moral decisions. We need to also be able to prove that we have done a good job. We need to be able to test whether the end behaviour we obtain is indeed ethical and correct with respect to some specification of correctness. For top-down agents testing and verification is made easier by knowing what their moral choices should be compared to. The top-down reasoning of the agent also allows itself to be formally verified [19, 18].

The motivation behind the bottom-up approach in building artificial agents draws from the observation that people typically make moral decisions without following a specific moral theory. We have a sense of right and wrong which we have developed over years of interactions with people, by observing how our decisions affect others and how we are affected by the decisions of others. The bottom-up approach aims to enable an artificial agent to learn little by little to discern right from wrong, emulating what people do.

To build a bottom up explicit ethical agent, one needs to figure out how to build an agent that learns to behave morally. In contrast, to build a bottom up implicit agent, one needs to figure out how to build an agent that given examples of right and wrong learns to identify moral choices correctly.

The advantages of the bottom-up approach are, clearly that one avoids the problem of choosing and implementing a specific moral theory. The approach is robust in the sense that it does not run the risk of running into a situation for which a moral decision cannot be made because the moral theory is under-specified. This robustness can also be a limitation of the bottom-up approach. As agents learn the moral behaviour they may end up learning something which we as humans do not recognise as moral. Examples are not hard to imagine, but we have already been able to witness some of them. Consider for example Tay[4]. Tay was a chatter-bot developed by Microsoft that was supposed to learn how to interact with people on Twitter, but was taken down after "learning" to post offensive tweets.

A limitation for explicit ethical bottom-up agent is also the dependence of its success on solving hard AI problems, just like the explicit ethical top-down agents. A specific limitation of the implicit ethical bottom-up agents is that they require reliable examples of right and wrong. This type of data is not readily available and it needs to be purposefully created. One challenge to creating this data set is that it would be an expensive undertaking. Another challenge is moral: who gets to supply the examples? How do we determine what are the representative examples of teachers of right and wrong for machines?

This shortage of examples has caused that deep learning, despite being the driver of much of the commercial success of recent AI [15], is virtually not used at all in machine ethics. Jiang et al [30] can perhaps be considered an exception. Jiang et al. use deep learning to build a question answering system that evaluates the morality of certain actions, however this system is not meant to be used by machines but by humans. Examples of bottom-up artificial agents either rely on symbolic learning, e.g., Anderson and Anderson [7] or on reinforcement learning [38, 9, 1]. The challenge of taking the reinforcement learning approach then becomes how to specify the objective function for the bottom-up artificial agent. This opens up again the problems of who gets to supply this information.

Lastly, compared to the top-down approach, the bottom-up approach does not lend itself as easy for testing and verification. What should the decisions of the artificial agent be compared to? Two moral agents might make two different choices in a moral decision making situation because they interpret the situation differently, and it is hard to claim that one choice is moral and the other is not.

The Ethical Turing test has been proposed [3] as a possible way to asses whether the artificial agent makes moral choices. This test would work as the original Turing test. An artificial moral agent would be given examples of decision problems and its answers will be compared to that of a moral philosopher or other expert in ethical decision making [7, 32]. Arguments have also been put forwards towards why the Ethical Turing test is not an adequate approach to testing if moral behavior in an artificial agent has been attained [10].

6 Beyond this tutorial

In this tutorial so far we had not discussed specific examples of artificial moral agents. This tutorial is not intended to be a systematic review of implemented machine ethics systems. Two such reviews exist and the reader can consult [49] and [37]. A very practical reason for

[4] `https://en.wikipedia.org/wiki/Tay_(bot)`

avoiding discussing implementations of artificial agents here is that these implementations vary vastly in the approaches they use and considerable background knowledge in various reasoning and learning methods would be necessary to understand the implementations. Throughout the text, however, numerous references are given to these specific systems and the interested reader can follow them and explore them for learning more.

It has to be mentioned that a considerable challenge to learn and conduct research in machine ethics is that research articles in machine ethics appear in a variety of AI venues, but also in volumes in engineering, decision theory, organisation theory and of course, philosophy.

References

1 David Abel, James MacGlashan, and Michael L Littman. Reinforcement learning as a framework for ethical decision making. In *Workshops at the thirtieth AAAI conference on artificial intelligence*, 2016.

2 Larry Alexander and Michael Moore. Deontological Ethics. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition, 2021.

3 Colin Allen, Gary Varner, and Jason Zinser. Prolegomena to any future artificial moral agent. *Journal of Experimental & Theoretical Artificial Intelligence*, 12(3):251–261, 2000. doi:10.1080/09528130050111428.

4 Michael Anderson and Susan Leigh Anderson. Machine ethics: Creating an ethical intelligent agent. *AI magazine*, 28(4), 2007.

5 Michael Anderson and Susan Leigh Anderson. The status of machine ethics: A report from the aaai symposium. *Minds Mach.*, 17(1):1–10, March 2007. doi:10.1007/s11023-007-9053-7.

6 Michael Anderson and Susan Leigh Anderson, editors. *Machine Ethics*. Cambridge University Press, 2011.

7 Michael Anderson and Susan Leigh Anderson. Geneth: A general ethical dilemma analyzer. In Carla E. Brodley and Peter Stone, editors, *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence, July 27 -31, 2014, Québec City, Québec, Canada*, pages 253–261. AAAI Press, 2014. URL: http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8308.

8 Susan Leigh Anderson. Asimov's "three laws of robotics" and machine metaethics. *AI Soc.*, 22(4):477–493, 2008. doi:10.1007/s00146-007-0094-5.

9 Stuart Armstrong. Motivated value selection for artificial agents. In *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

10 Thomas Arnold and Matthias Scheutz. Against the moral turing test: Accountable design and the moral reasoning of autonomous systems. *Ethics and Inf. Technol.*, 18(2):103–115, June 2016. doi:10.1007/s10676-016-9389-x.

11 Seth D. Baum. Social choice ethics in artificial intelligence. *AI Soc.*, 35(1):165–176, 2020. doi:10.1007/s00146-017-0760-1.

12 Tom L. Beauchamp and James F. Childress. *Principles of Biomedical Ethics*. Oxford University Press, USA, 2001.

13 Richard E. Bellman. *An Introduction to Artificial Intelligence: Can Computers Think?* Boyd & Fraser Publishing Company, 1978.

14 Oliver Bendel, Kevin Schwegler, and Bradley Richards. Towards kant machines. In *2017 AAAI Spring Symposia, Stanford University, Palo Alto, California, USA, March 27-29, 2017*. AAAI Press, 2017. URL: http://aaai.org/ocs/index.php/SSS/SSS17/paper/view/15278.

15 Yoshua Bengio, Yann Lecun, and Geoffrey Hinton. Deep learning for AI. *Communications of the ACM*, 64(7):58–65, June 2021. doi:10.1145/3448250.

16 Bartosz Brożek and Bartosz Janik. Can artificial intelligences be moral agents? *New Ideas in Psychology*, 54:101–106, 2019. doi:10.1016/j.newideapsych.2018.12.002.

17 Alexandra Chouldechova and Aaron Roth. A snapshot of the frontiers of fairness in machine learning. *Commun. ACM*, 63(5):82–89, 2020. doi:10.1145/3376898.

18 Louise A. Dennis, Martin Mose Bentzen, Felix Lindner, and Michael Fisher. Verifiable machine ethics in changing contexts. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 11470–11478. AAAI Press, 2021. URL: https://ojs.aaai.org/index.php/AAAI/article/view/17366.

19 Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster. Formal verification of ethical choices in autonomous systems. *Robotics Auton. Syst.*, 77:1–14, 2016. doi:10.1016/j.robot.2015.11.012.

20 Nicholas Diakopoulos. Transparency. In Markus D. Dubber, Frank Pasquale, and Sunit Das, editors, *The Oxford Handbook of Ethics of AI*. Oxford University Press, July 2020. doi:10.1093/oxfordhb/9780190067397.013.11.

21 Franz Dietrich and Christian List. What matters and how it matters: a choice-theoretic representation of moral theories. *Philosophical Review*, 126(4):421–479, 2017.

22 Virginia Dignum. *Responsible Artificial Intelligence - How to Develop and Use AI in a Responsible Way*. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, 2019. doi:10.1007/978-3-030-30371-6.

23 Sjur Dyrkolbotn, Truls Pedersen, and Marija Slavkovik. On the distinction between implicit and explicit ethical agency. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '18, pages 74–80, New York, NY, USA, 2018. Association for Computing Machinery. doi:10.1145/3278721.3278769.

24 Amitai Etzioni and Oren Etzioni. Incorporating ethics into artificial intelligence. *The Journal of Ethics*, 21:403–418, 2017.

25 James Fieser. Ethics. In Michael Boylan, editor, *Internet Encyclopedia of Philosophy*. ISSN 2161-0002, 2021.

26 Dov Gabbay, John Horty, and Xavier Parent. *Handbook of Deontic Logic and Normative System*. College Publications, UK, 2013.

27 James Gips. Toward the ethical robot. In Kenneth M. Ford, C. Glymour, and Patrick Hayes, editors, *Android Epistemology*, pages 243–252. MIT Press, USA, 1994.

28 David Gunning and David Aha. Darpa's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2):44–58, June 2019. doi:10.1609/aimag.v40i2.2850.

29 R. M. Hare. *Community and Communication*, pages 109–115. Macmillan Education UK, London, 1972. doi:10.1007/978-1-349-00955-8_9.

30 Liwei Jiang, Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. Delphi: Towards machine ethics and norms. *CoRR*, abs/2110.07574, 2021. arXiv:2110.07574.

31 Gregory E Kersten and Stan Szpakowicz. Decision making and decision aiding: defining the process, it representations, and support. *Group Decision and Negotiation*, 3(2):237–261, 1994.

32 Hyeongjoo Kim and Sunyong Byun. Designing and Applying a Moral Turing Test. *Advances in Science, Technology and Engineering Systems Journal*, 6(2):93–98, 2021. doi:10.25046/aj060212.

33 Bertram F. Malle, Paul Bello, and Matthias Scheutz. Requirements for an artificial agent with norm competence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, pages 21–27, New York, NY, USA, 2019. Association for Computing Machinery. doi:10.1145/3306618.3314252.

34 Tim Miller. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, 267:1–38, 2019. doi:10.1016/j.artint.2018.07.007.

35 James Moor. The dartmouth college artificial intelligence conference: The next fifty years. *AI Magazine*, 27(4):87, December 2006. doi:10.1609/aimag.v27i4.1911.

36 James H. Moor. The nature, importance, and difficulty of machine ethics. *IEEE Intelligent Systems*, 21(4):18–21, July 2006. doi:10.1109/MIS.2006.80.

37 Vivek Nallur. Landscape of machine implemented ethics. *Sci. Eng. Ethics*, 26(5):2381–2399, 2020. doi:10.1007/s11948-020-00236-y.

38 Ritesh Noothigattu, Djallel Bouneffouf, Nicholas Mattei, Rachita Chandra, Piyush Madan, Kush R. Varshney, Murray Campbell, Moninder Singh, and Francesca Rossi. Teaching ai agents ethical values using reinforcement learning and policy orchestration. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 6377–6381. International Joint Conferences on Artificial Intelligence Organization, July 2019. doi:10.24963/ijcai.2019/891.

39 Ritesh Noothigattu, Snehalkumar (Neil) S. Gaikwad, Edmond Awad, Sohan Dsouza, Iyad Rahwan, Pradeep Ravikumar, and Ariel D. Procaccia. A voting-based system for ethical decision making. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 1587–1594. AAAI Press, 2018. URL: https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17052.

40 Raja Parasuraman, Tom .B. Sheridan, and Christopher D. Wickens. A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 30(3):286–297, 2000. doi:10.1109/3468.844354.

41 Martin Peterson. *An Introduction to Decision Theory*. Cambridge Introductions to Philosophy. Cambridge University Press, 2 edition, 2017. doi:10.1017/9781316585061.

42 Rosalind W. Picard. *Affective Computing*. MIT Press, 1997.

43 Thomas M. Powers. Prospects for a kantian machine. In Michael Anderson and Susan LeighEditors Anderson, editors, *Machine Ethics*, pages 464–475. Cambridge University Press, 2011. doi:10.1017/CBO9780511978036.031.

44 Iyad Rahwan. Society-in-the-loop: programming the algorithmic social contract. *Ethics and Information Technology*, 20(1):5–14, March 2018. doi:10.1007/s10676-017-9430-8.

45 Steward Russell and Peter Norvig. *Artificial Intelligence: A Modern Approach*. Pearson Education, 4 edition, 2020.

46 Amartya Sen. Choice, orderings and morality. In Stephan Körner, editor, *Practical Reason*, pages 54–67. Camalot Press, Oxford, 1974.

47 Amartya K. Sen. Rational fools: A critique of the behavioral foundations of economic theory. *Philosophy & Public Affairs*, 6(4):317–344, 1977. URL: http://www.jstor.org/stable/2264946.

48 Jakob Stenseke. Artificial virtuous agents: from theory to machine implementation. *AI & SOCIETY*, 2021. doi:10.1007/s00146-021-01325-7.

49 Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: A survey. *CoRR*, abs/2001.07573, 2020. arXiv:2001.07573.

50 Lewis Vaughn. *Beginning Ethics: An Introduction to Moral Philosophy*. W. W. Norton & Company, New York City, 2014.

51 Wendell Wallach and Colin Allen. *Moral Machines: Teaching Robots Right from Wrong*. Oxford University Press, Inc., USA, 2008.

52 Wendell Wallach, Colin Allen, and Iva Smit. Machine morality: bottom-up and top-down approaches for modelling human moral faculties. *AI & SOCIETY*, 22(4):565–582, 2008. doi:10.1007/s00146-007-0099-0.

53 Maranke Wieringa. What to account for when accounting for algorithms: A systematic literature review on algorithmic accountability. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, pages 1–18, New York, NY, USA, 2020. Association for Computing Machinery. doi:10.1145/3351095.3372833.

54 Alan F. T. Winfield, Serena Booth, Louise A. Dennis, Takashi Egawa, Helen Hastie, Naomi Jacobs, Roderick I. Muttram, Joanna I. Olszewska, Fahimeh Rajabiyazdi, Andreas Theodorou, Mark A. Underwood, Robert H. Wortham, and Eleanor Watson. Ieee p7001: A proposed standard on transparency. *Frontiers in Robotics and AI*, 8:225, 2021. doi:10.3389/frobt.2021.665729.