Engineering moral machines

Engineering Moral Machines

Michael Fisher, Christian List, Marija Slavkovik, and Alan Winfield[1]
mfisher@liverpool.ac.uk, c.list@lse.ac.uk, marija.slavkovik@uib.no, alan.winfield@uwe.ac.uk

Abstract

This article provides a short report on a recent Dagstuhl Seminar on "Engineering Moral Agents". Imbuing robots and autonomous systems with ethical norms and values is an increasingly urgent challenge, given rapid developments in, for example, driverless cars, unmanned air vehicles (drones), and care assistant robots. Seminar participants discussed two immediate problems. A challenge for philosophical research is the formalisation of ethics in a format that lends itself to machine implementation; a challenge for computer science and robotics is the actual implementation of moral reasoning and conduct in autonomous systems. This article reports on these two challenges.

Introduction

Machines and software with limited autonomy have existed in our society for many years. To guarantee the safety and well-being of their users and others, two strategies have traditionally been employed. Either the systems in question are used only in strictly controlled environments, as in the case of automated trains and factory robotic arms, or they are designed to have very limited abilities to manipulate their environments, as in the case of robotic floor cleaners. However, systems with increased autonomy and intentionality, for short "AI systems", are an imminent reality. Prominent examples are driverless vehicles, robotic surgical systems, and care assistant robots. There is also increasing automation in civil aviation and in military drones.

AI systems require sufficient physical capabilities to interact with their environments in the intended ways, yet they also often share their operation space with people. This means that many of the traditional solutions for ensuring operational safety, legality, and compliance with moral norms are no longer applicable. A number of concerns arise once advanced AI systems and people share the same environment. These concerns include, but are not limited to:

- ensuring the safety of the people who share their space with AI systems;
- developing operational standards and certification methods for such systems;
- determining who is legally responsible for their operation, and ensuring compliance with legal and moral norms;
- defining the scope of responsibility for any harms caused;
- designing the systems to recognise, and correctly to respond to, moral decision problems in time-critical situations; and
- preventing abuse of these systems by people for illegal or immoral purposes.

Many countries are beginning to develop legal frameworks and industrial standards for open-market availability and widespread use of certain AI systems, such as unmanned aerial vehicles

[1] The authors are very grateful to all the participants of the Dagstuhl Seminar on "Engineering Moral Agents" for their contributions to the discussion and to John Horty, Marcus Pivato, and Kai Spiekermann for helpful comments on a draft of this article.

and driverless cars. As is commonly recognised, the problems involved are not just engineering problems, but conceptual and philosophical ones.

Moral philosophy goes back to antiquity, long pre-dating even the fictional consideration of AI systems. For this reason, the vast body of thought in moral philosophy is premised on the idea that moral agents are human. Relatively little consideration has been given to the possibility that systems other than individual human beings might qualify as moral agents (exceptions can be found in discussions of group agency, e.g., in French 1984 and List and Pettit 2011). AI systems differ fundamentally from humans. Their hardware and software are very different from human biology and psychology, and human beings have many characteristics that AI systems typically lack: for example, they are conscious, mortal, emotional, dependent on society, born and raised by other people, and trained and motivated by their peers. Therefore traditional lessons from moral philosophy may not be directly transferable to artificial agents.

Moral issues arise, not only when AI systems directly influence and affect people's lives, but also when they aid human decision-making or indirectly facilitate human activities. The relevant area of research has become known as "machine ethics" or "artificial morality". It draws on many disciplines, such as robotics, computer science, philosophy, psychology, law, and economics.

The Dagstuhl Seminar 16222 on "Engineering Moral Agents", held in May 2016, brought together researchers from several disciplines to review the current state of the field and to identify key challenges for future research. Much of the discussion revolved around questions concerning how to develop a moral framework for "intelligent" machines. How can we implement moral reasoning in AI systems? How might we build regulatory structures that address (un)ethical machine behaviour? What are the wider societal, legal, and economic implications of introducing AI systems into our society?

The seminar participants compared two leading approaches to engineering moral machines, the so-called "constraint-based" and "training approaches", and discussed two immediate challenges. A challenge for philosophical research is the formalisation of ethics in a format that subsequently lends itself to machine implementation; a challenge for computer science and robotics is the actual implementation of moral reasoning and conduct in autonomous systems. In this article, we briefly report on the discussion.

Two approaches to engineering moral machines

It is generally recognised that there are two core approaches to engineering moral machines (Allen et al. 2005):

- a constraint-based approach: explicitly constraining the actions of an AI system in accordance with moral norms; and

- a training approach: training the AI system to recognise and correctly respond to morally challenging situations.

These are sometimes also called "top-down" and "bottom-up". In addition, there are also hybrid approaches. Let us briefly discuss these approaches in turn.

The constraint-based approach involves identifying a set of rules or principles that an AI system has to follow and then implementing them, so as to pre-check or constrain the system's actions. Isaac Asimov's famous laws of robotics are an example of such rules. Asimov's laws require that, first, robotic systems must not harm humans or allow them to be harmed; second, they must obey human orders provided this does not conflict with the first law; and third, they must

protect themselves provided this does not conflict with the first two laws. In reality, of course, these laws are just a very simple, illustrative starting point, and more complex and nuanced sets of norms are needed for genuine action guidance. What if a machine is faced with a choice between avoiding a grave harm to one person and avoiding a lesser harm to several people, for example?

In principle, some moral theories, such as utilitarianism and Kantian deontology, may be amenable for devising constraints on action, provided those theories can be suitably formalised (cf. Allen 2005). Utilitarian theories, in particular, have been formalized in disciplines such as decision theory and social choice theory (e.g., Broome 1991). Arguably, however, common-sense morality is not utilitarian, but involves deontological principles and various heuristics. Furthermore, moral rules and principles are often vague and context-dependent, and there can be conflicts between them. Both formalisation and conflict resolution remain significant challenges, even within moral philosophy. It is fair to say that we currently have no complete formalization of common-sense morality.

A training approach involves applying techniques such as machine learning to "educate" an AI system to recognise morally challenging situations and to resolve conflicts, much as human beings are educated by their carers and community to become moral agents. Until recently, limitations in computational power have restricted the scope of such an approach, but advances in computing, especially in processing large data sets, make it increasingly feasible. A training approach avoids some of the problems of a constraint-based approach. In particular, we do not require a completely formalized moral theory for it. A sufficiently rich training database of illustrative moral decision problems, with the appropriate target judgments, is sufficient for the approach to get off the ground. The machine's acquisition of morality would be much like the way in which a Bayesian learning system can be trained to recognize cancer cells or other salient patterns in medical images.

But the approach comes at a cost, since training is slow, resource-intensive, error-prone, and may – in the extreme case – have to be done anew for each different AI system. Moreover, we would require a compelling database of examples of what it means to behave ethically or unethically, and there is much disagreement on what the correct moral judgments would be, even among moral philosophers. Finally, an AI system that learns its morality through big data may not be able to *explain* why the actions it chooses are moral. The ability to explain one's actions is often considered a crucial feature of moral agency.

Much work in machine ethics, up to now, has been exploratory, describing and debating the feasibility of artificial morality, its implementation, and the relevant social impact. Some constraint-based and training systems have been developed, typically to serve as a proof of concept. An example of a training approach is discussed in Anderson and Anderson (2014), while examples of constraint-based approaches are discussed in Arkin et al. (2012), Winfield et al. (2014), and Dennis et al. (2016).

A clear advantage of a constraint-based approach – especially if it involves the symbolic representation of moral reasoning – is the possibility (at least in principle) of using formal verification to test that the reasoning works as intended. If a training approach is used, on the other hand, the training should ideally take place before the autonomous system is deployed in practice, and the system's moral behaviour should be somehow tested. A possible path to certifying trained systems is considered in Anderson et al. (2016). However, regulatory bodies and other public institutions have yet to determine what exactly characterises an autonomous system that is "safe to deploy".

Against this background, the seminar participants at Dagstuhl split into two discussion groups. The first considered the formalisation of moral reasoning, the second its implementation.

Formalising moral reasoning

To illustrate what sorts of questions moral reasoning must generally address, consider two stylized examples of moral decision problems. The first comes from Scanlon (1998, p. 235):

> "Jones has suffered an accident in the transmitter room of a television station. Electrical equipment has fallen on his arm, and we cannot rescue him without turning off the transmitter for fifteen minutes. A World Cup match is in progress, watched by many people, and it will not be over for an hour. Jones's injury will not get any worse if we wait, but his hand has been mashed and he is receiving extremely painful electrical shocks. Should we rescue him now or wait until the match is over? Does the right thing to do depend on how many people are watching... ?"

The second example is the well-known trolley problem, introduced by Foot (1967), which we here summarise as follows:

> A run-away trolley races down a track. At the end of the track, there are five people, who will be run over by the trolley and killed if the trolley is not diverted to a side-track. At the end of the side-track, however, there is one person, who will be run over and killed if the trolley is diverted. You are in control of a switch to determine whether or not to divert the trolley onto the side-track. Should you divert the trolley?

In each of these cases, we – human beings – have certain moral intuitions as to what the right action is. Sometimes we have conflicting intuitions, and different moral principles yield different verdicts. Moral theories are an attempt to systematise our moral intuitions, in order to deduce them from some underlying principles and to explain them. The challenge for machine ethics is to encode moral theories in a machine-implementable way. This often requires formalization, at least if we opt for a constraint-based approach to engineering moral machines.

The discussion group considered several approaches to the formalization of moral theories. The first set of approaches uses logic, such as deontic logic or default logic, to represent moral reasoning explicitly. Candidate formalisms can be found in classical deontic logic, but also in the recent work of several of our seminar participants. John Horty, for instance, uses a version of default logic to represent legal and moral reasoning (Horty 2012). Marek Sergot applies variants of deontic logic and STIT ("see to it that") logic to represent normative relations between agents (e.g., Sergot 2013). Both Horty and Sergot presented some of their ideas to the group. Horty talked about developing a computational theory of moral reasoning based on his representation of moral rules in default logic, and Sergot talked about value-based argumentation and prioritised defeasible conditional imperatives.

A logical formalization of a moral theory is symbolic and thereby lends itself, in principle, to verification and validation: formally proving that a moral system has the intended properties. Moreover, many of the leading logical formalisms can accommodate several competing moral theories, so that the formalisms do not by themselves dictate the resulting moral judgments. In particular, unlike some classical decision-theoretic approaches, they are not automatically committed to some version of consequentialism, but can capture deontological theories too. Insofar as common-sense morality is arguably not consequentialist, this approach clearly holds some promise.

A second set of approaches comes from decision theory. Here, the idea is to apply insights from microeconomics to the formalisation of moral theories. Formalizations of utilitarianism and

other consequentialist theories, for instance, explicitly introduce utility or welfare functions for all affected subjects of moral concern – for instance, all people that might be affected by a decision – and then represent moral reasoning as an optimization problem: the goal may be, for instance, to maximize expected total utility. One of our participants, Marcus Pivato, gave an overview of such formalizations.

More generally, we may ask whether even those moral theories that are not overtly consequentialist can be translated into a consequentialist format. A moral theory is said to be "consequentializable" if its action-guiding recommendations can be represented by a choice function that is induced by some linear ordering (a "betterness ordering") over all actions under consideration (e.g., Brown 2011). There is considerable debate in moral philosophy about whether all moral theories can be consequentialized, at least in principle. The discussion group concluded – in agreement with a number of philosophers – that consequentialization has its formal limits. We can consequentialize some conventionally non-consequentialist theories only at the cost of stretching or redefining the notion of "consequence" (for discussions, see, e.g., Brown 2011, Dietrich and List 2016). If we are willing to build various contextual features into the notion of "consequence", then "consequentialization" becomes vacuously possible, but it will no longer be very useful for the purpose of encoding moral theories in a machine-implementable way.

A third approach was presented by Christian List, based on joint work with Franz Dietrich. This approach is an attempt to represent a large class of moral theories in a canonical format, without "consequentializing" them in a potentially trivialising manner. Specifically, Dietrich and List (2016) offer a "reason-based" formalisation of moral theories. They encode the action-guiding content of a moral theory in terms of a choice function (here they share the starting point of the standard decision-theoretic approach), which they interpret as a "rightness" function. Formally, this is a function that assigns to each set of feasible actions or options the subset of morally permissible ones. Instead of consequentializing this rightness function, they show that any rightness function within a large class can be represented in terms of two parameters: (i) a specification of which properties of the options are normatively relevant in any given context, and (ii) a betterness relation over sets of properties. Importantly, the normatively relevant properties need not be restricted to "consequence properties" alone, but can also include "relational properties", i.e., properties specifying how options relate to the context of choice. A relational property might be, for instance, whether a given option satisfies some context-specific moral norm.

Reason-based representations provide a taxonomy of moral theories, as theories can be classified in terms of parameters (i) and (ii) above. We may ask: are the same properties normatively relevant in all contexts? If so, the theory is universalistic; if not, the theory is relativistic. Further, are the normatively relevant properties restricted to "consequence properties"? If so, the theory is consequentialist; if not, it is non-consequentialist.

The discussion group recognised that the same action-guiding recommendations can often be systematised by competing moral theories (as argued, e.g., by Broome 2004, ch. 3, and Dietrich and List 2016). This is a consequence of the fact that moral theories specify not only *how* we ought to act, but also *why* we ought to act in that way. Different answers to the "why" question may be compatible with the same answer to the "how" question. An interesting issue, therefore, is whether moral machines need to get only the "how" question right, or whether the "why" question matters for them as well.

The discussion group further acknowledged the need to take resource-boundedness into account when we formalise moral theories. Joseph Halpern, in particular, explained this aspect (see also Halpern et al. 2014). If our goal is to arrive at machine-implementable moral theories, we cannot

presuppose complete information and unlimited computational capacities. While the idea of bounded rationality has received much attention in psychology and economics, there is no equally well-developed analogue of this idea for morality: a notion of "resource-bounded morality". There is some work on "ideal versus non-ideal theory" in moral philosophy, but this is primarily concerned with the morality of institutions and institutional design, and less with individual agents whose agentive capacities are limited. One interesting question is whether, under informational and computational constraints, rule-based, deontological, or virtue-ethical approaches might outperform consequentialist approaches, for which optimization is central (cf. Slote 1989). That said, we can also define versions of utilitarianism that are based on the idea of *constrained* optimisation. Similarly, some versions of *rule utilitarianism* (or more generally, *rule consequentialism*) are attempts to reconcile consequentialist moral philosophy with computational and informational constraints.

Finally, the discussion group noted that, at present, moral reasoning focuses – rightly – on human beings and other sentient animals (perhaps also the environment) as the ultimate units of moral concern. This raises the question of whether, and to what extent, the (still hypothetical) development of AI consciousness might require a more significant rethinking of our anthropocentric moral codes.

Implementing moral reasoning

The key challenge for computer scientists, roboticists, and engineers is not so much the question of which moral theory to implement, but rather the actual implementation of moral reasoning in AI systems. There are advantages and disadvantages of both standard approaches to implementing moral reasoning: top-down and bottom-up. In a top-down approach, one begins with a well-defined task or objective that is to be solved by the system. The system is then designed to fulfil these requirements in the given environment or on the given data. In a bottom-up approach, the environment or data is the starting point. The system then uses some form of automated learning in order to detect patterns in that environment or data and to perform the desired task on the basis of the detected patterns. This approach typically requires large amounts of data as input: a training database. A hybrid approach is also possible, but it is less clear what its advantages and disadvantages would be.

The discussion group considered issues of specification and verification with respect to both approaches, which in turn raised issues of transparency and accountability. The problem of (formal) verification is to prove that an autonomous system's actions are within the bounds of moral behaviour for the context in which it operates. The issue of transparency, as with much other complex machinery, is the issue of the level of detail of operation that will be made accessible to different concerned entities such as the end user, the manufacturer, licensed maintenance personnel, government regulatory bodies, and so on.

When a machine is in a position to cause harm to, or even death of, numerous people, such as the autopilot of a passenger airplane, certain safety standards are required. An autopilot is considered safe if it operates without causing an accident in a certain "very high" number of cases. It seems evident that such safety requirements will need to be specified for AI systems capable of making decision in situations where moral considerations are relevant. The question of how safe is "safe enough" needs to be further discussed in this context. Taking passenger aircraft as an exemplar, the discussion group agreed on the need for some classes of AI systems – driverless cars for instance – to be equipped with "ethical black boxes": devices that will allow the internal moral decision-making processes to be recorded for later review during, for example, accident investigations.

The group discussed possible effects that a moral reasoning machine can have on society. By implementing one moral code rather than another, the manufacturer of a device may be implicitly imposing one culture's morality on another culture, which might have different values than the manufacturer's. In addition, introducing machines capable of moral reasoning into a society may have unpredictable effects on that society and on how people behave towards such machines. The behaviour of the machines may not cause any physical harm, but it may set in motion events that inevitably lead to unintended social or psychological harms. These risks must be taken into consideration when AI systems are designed and deployed.

The group also recognized the importance of protecting the operation of AI systems from malicious or mischievous influence by users and society, which the group termed "the dark side" of moral machines. Each of the approaches to implementing moral reasoning is susceptible to different kinds of vulnerabilities, which must be taken into account.

In the final closing discussion, Kai Spiekermann made an important methodological remark about the implementation of moral reasoning in AI systems, related to the earlier discussion of the formalization of moral theories. He noted that it is sometimes assumed that the implementation of moral reasoning in machines is possible only once we have found a way of completely formalizing common-sense morality. But it may well be that no such formalization can ever be achieved. As Spiekermann pointed out, we also do not expect *human* moral agents to have a completely specified (let alone formalized) moral code at their disposal. To the contrary, while human moral agents have clear intuitions about, and agree on, clear-cut cases, such as the importance of avoiding unnecessary suffering, they often lack settled judgments about more complicated cases and passionately argue about them. Examples of such cases are the trade-offs between different values or the resolution of moral dilemmas. It appears that having a complete moral code is not what we normally require for moral agency, and, by extension, it may be implausible to require machines to have a complete moral code. The properties that appear to make someone (or something) a moral agent are the following: they include the ability to reason and to justify moral choices, and perhaps they also include certain complex psychological abilities, such as the ability to empathize. A promising avenue for engineering artificial moral agents may be to work towards building systems that have these abilities, rather than requiring the implementation of a completely specified moral code.

Concluding remarks

The moral behaviour of machines is a topic of growing urgency, particularly with the prospect of increasingly advanced AI systems being introduced into our societies in the coming years. It is essential to ensure, first, that the reasoning implemented in those machines is designed by experts across robotics, computer science, philosophy, psychology, law, and economics who have a sufficiently deep understanding of the relevant issues, and second, that it is scrutinized by a well-informed public debate. The design of moral machines must not be left exclusively to companies and manufactures without critical expert input or a transparent process of public scrutiny. Efforts such as this seminar, endeavouring to discuss and clarify the challenges of machine ethics, are vital to ensuring not only that AI systems are reliable, but also that the public will trust them enough to rely on them. This is an ongoing process, and the relevant discussions, across all disciplines and across all strands of society, must continue.

References

Colin Allen, Iva Smit, and Wendell Wallach: Artificial morality: Top-down, bottom-up, and hybrid approaches. *Ethics and Information Technology* 7: 149-155, 2005.

Michael Anderson, Susan Leigh Anderson, and Vincent Berenz: Ensuring Ethical Behaviour from Autonomous Systems. *Proc. AAAI Workshop: Artificial Intelligence Applied to Assistive Technologies and Smart Environments*, 2016. http://www.aaai.org/ocs/index.php/WS/AAAIW16/paper/view/12555

Michael Anderson and Susan Leigh Anderson: GenEth: A General Ethical Dilemma Analyzer. *Proc. AAAI*, pp. 253-261, 2014. http://www.aaai.org/ocs/index.php/AAAI/AAAI14/paper/view/8308

Ronald C. Arkin, Patrick Ulam, and Alan R. Wagner: Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proc. IEEE* 100 (3): 571–589. http://ieeexplore.ieee.org/document/6099675/?arnumber=6099675

John Broome: *Weighing Goods: Equality, Uncertainty and Time*. Oxford (Blackwell), 1991.

John Broome: *Weighing Lives*, Oxford (Oxford University Press), 2004.

Campbell Brown: Consequentialize This. *Ethics* 121(4): 749-771, 2011. www.jstor.org/stable/10.1086/660696

George C. Christie: *The Notion of an Ideal Audience in Legal Argument*. Dordrecht (Kluwer Academic Publishers), 2000. http://dx.doi.org/10.1007/978-94-015-9520-9

Jules L. Coleman: *Risks and Wrongs*. Oxford (Oxford University Press), 2002. http://dx.doi.org/10.1093/acprof:oso/9780199253616.001.0001

Louise A. Dennis, Michael Fisher, Marija Slavkovik, and Matt Webster: Formal Verification of Ethical Choices in Autonomous Systems. *Robotics and Autonomous Systems* 77: 1-14, 2016. http://dx.doi.org/10.1016/j.robot.2015.11.012

Franz Dietrich and Christian List: What matters and how it matters: A choice-theoretic representation of moral theories. Working paper, *London School of Economics*, 2016. http://personal.lse.ac.uk/list/PDF-files/WhatMatters.pdf

Philippa Foot: The Problem of Abortion and the Doctrine of the Double Effect in Virtues and Vices. *Oxford Review*, Number 5, 1967. http://philpapers.org/archive/FOOTPO-2.pdf

Peter A. French: *Collective and Corporate Responsibility*. New York (Columbia University Press), 1984.

Joseph Y. Halpern, Rafael Pass, and Lior Seeman: Decision Theory with Resource-Bounded Agents. *Topics in Cognitive Science* 6: 245-257, 2014. doi:10.1111/tops.12088

John Horty: *Reasons as Defaults*. New York / Oxford (Oxford University Press), 2012.

Christian List and Philip Pettit: *Group Agency: The Design, Possibility, and Status of Corporate Agents*. Oxford (Oxford University Press), 2011.

Thomas M. Scanlon: *What we owe to each other*. Cambridge, Massachusetts: Cambridge, MA (Belknap Press of Harvard University Press), 1998. http://www.hup.harvard.edu/catalog.php?isbn=9780674004238&content=reviews

Marek Sergot: Normative Positions. In Dov Gabbay, John Horty, Ron van der Meyden, Xavier Parent, and Leendvert van der Torre (eds.), *Handbook of Deontic Logic and Normative Systems*, Chapter 5, pp. 353-406, London (College Publications), 2013.
Michael Slote: *Beyond Optimizing: A Study of Rational Choice*. Cambridge, MA (Harvard University Press).

Alan Winfield, C. Blum, and W. Liu: Towards an ethical robot: Internal models, consequences and ethical action selection. In: M. Mistry, Aleš Leonardis, M. Witkowski, and C. Melhuish (eds.), Advances in Autonomous Robotics Systems, LNCS Vol 8717, pp. 85-96, Heidelberg (Springer), 2014.