Should I kill or rather not? [1]

Luis Moniz Pereira [2]

Robots are already among us: They build our cars and vacuum our apartments. Why do these machines need a sense of right and wrong?

Their tasks will change: They will work a lot closer with humans, and they'll have more autonomy, for example as caretakers for the elderly. Imagine a robot in a nursing home. It's helping elderly people with eating and grooming, and it's handing out medicine. One morning a resident asks the robot for painkillers, because he has a terrible headache. The robot is allowed to hand out pills only with the approval of a doctor. But none of the doctors are available. Will the robot let the resident suffer, or will it make an exception? Its decision depends on the way we program it.

The thought that a robot will take such decisions for us is a little eerie.

I think the problem is not that machines will take over. I think it's that we're giving too much power to simplistic machines: Machines that take decisions based on statistics. They can neither consider the individual circumstances of each case, nor can they justify their actions.

You say that to teach a robot morality, we need to know what we as humans consider right or wrong. How much do we know about our own moral principles?

---

[1] This is a joint synopsis, in English, of 3 interviews on the subject of "Machine Ethics" given by the author in 2018. Nora Saager, a science journalist for the German "P. M. Magazine", conducted one of them. The original, in German, came out in its February 2018 issue. Another was conducted by journalist Pedro Lucas for the feature "Um Café Com..." of "Men's Health" magazine. The original, in Portuguese, was published in its January 2018 issue. Journalist Virgílio Azevedo, for his regular feature "O Futuro do Futuro" in the weekly "Expresso", conducted the third. The original, in Portuguese, was published on 28 April 2018. I thank the 3 journalists for permission to utilize my fusion and English translation of selected parts of the above-mentioned material.

[2] NOVA-LINCS research centre and Departamento de Informática, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Portugal.
Web: http://userweb.fct.unl.pt/~lmp/ Email: lmp@fct.unl.pt

Neither computer scientists nor sociologists know enough about human morality. Nobody does. One thing seems certain: Morality evolved. We're a gregarious species, so we need rules for living together. We are born with the ability to learn moral behaviour – much like we're born with the ability to learn a language. 95 per cent of all moral decisions are taken by reflex. It's only in complex situations that we need to think things through or even suppress our first impulse.

So we're deciding intuitively, without knowing why?

At least, most people have difficulties explaining why they decided one way or the other. And that's a problem: We don't know about the basics of morality to program it. Even ethicists disagree on how to act in certain moral dilemmas. They disagree on what constitutes good moral reasoning. There are different schools of thought.

Wouldn't a machine be perfectly suited to calculate which decision yields the greatest benefit for all involved?

But how would you do it? And which information is needed? Imagine a situation where you need to decide who lives and who dies. Is it better to save a doctor, who in turn might go on to save many more lives? Or do you save a young person who has his whole life ahead of him? There is no simple, universal morality that everyone can agree on.

Sounds like it's almost a hopeless mission to program a machine with fixed moral rules.

We're still at the very beginning. We should start with clearly defined norms for specific settings: for hospitals, for childcare, for nursing homes. The most detailed and widely subscribed rules that exist are probably and unfortunately those for warfare.

You wrote a program where a robot has to save a princess and make moral decisions along the way – not exactly an everyday situation. What's the intention behind the simulation?

The program shows how to combine different moral approaches. The number of rules for the robot's behaviour increases with every step. Sometimes, they are

contradictory. In that case, the robot has to set priorities. It is also able to imagine different scenarios – a very important ability for making moral decisions.

Does the ideal machine act as human-like as possible?

One day, intelligent machines will live alongside humans. And we will accept them only if their morality is very similar to ours. Assuming we're sending robots to Mars, with no humans to accompany them: Then we can give them morals that are custom-made for this environment.

Why would they need to act morally in space, far away from human company?

They'll probably have different manufacturers – so we can't assume they all use the same software. Still, they need to cooperate. We're not sending robots to Mars only to watch them destroy each other. The same goes for our home planet: When machines from different manufacturers work together, for example as guards in a shopping mall, there's always a danger of competition. There's also a risk that a robot is programmed with sinister intentions. Maybe it wasn't sent by a security firm, but by thieves? One purpose of morality is to detect cheaters and freeloaders.

Some researchers promote machine learning in morality: Using artificial intelligence, computers can comb through large amounts of data.

Right now, there's a lot of hype surrounding this method. Machine learning works very well in some fields – for example those related to perception. But it won't work for morality.

Why not?

I'll give you an example. A while ago, there was an accident with a self-driving car in Tulsa, Arizona. The car went through an intersection when the lights had turned yellow, and it crashed into an oncoming car, which was turning left. Pretty bad stuff. What had happened? The self-driving car had followed the rules. But it hadn't considered that other drivers might not do so. That's the problem: The software needs accidents to learn. Yet, accidents shouldn't be happening! That's why you need preventive rules. Another problem is: Algorithms can't explain their decisions. If you asked them "Why did you do this," the answer would be: "I was subjected to a

number of simulations and analysed them using statistics." But if I live alongside a robot, I want it to know the reason for its decisions.

Self-learning machines might even act completely unpredictably. We don't know which conclusions they draw from the data they've been fed.

Exactly. And that raises another point: How do you test moral software? Control software for trains or aircraft is written in a certain way, to make sure some things cannot happen. Now, if you write moral code, you need to include basic rules that will never be broken. For example: If deaths are absolutely unavoidable, make sure the very least number of people is killed. But how can I prove that my program will always make the right decision, no matter what moral dilemma it's confronted with?

Who should decide which moral values we teach machines? Do we need an international expert panel?

Whoever decides: The result needs to be written in law, which is monitored and enforced by national governments and international contracts.

You're not just trying to teach machines morality – you're also using computers to better understand human behaviour. How do you do that?

To investigate moral behaviour in groups, I let virtual agents with different approaches compete in a simple game. The most successful strategy will become more widespread, and it will be passed on to following generations. In one of my recent studies, I showed that guilt promotes cooperation. If cheaters feel guilty and show remorse, they will benefit more from future interactions. Others will copy his behaviour, and he might have offspring who will also feel guilty. Such a mathematical model shows: There's a reason why guilt developed and spread through society. It also shows: If we want to program moral machines, we should give them a sense of guilt.

You believe we could also simulate the effects of new legislation before we pass them.

We know for example that we'll need legislation and jurisdiction for robots. And those need a moral basis. That prompts a lot of questions, such as: To what degree are robots responsible for their actions? If lawmakers simulate certain cases, they can try

out different moral guidelines and assess the outcome. The computer would be a tool to experiment with moral principles.

Are we well on track to build a robot that acts decently, or is that still a long way off?

Programming morality is a very complex problem; it has many dimensions. We're just starting to understand the challenges. It's like exploring a new continent. The Portuguese have done this many times in the past – maybe that's one of the reasons why I'm exploring this uncharted territory, about which you can find more in our book, L. M. Pereira and A. Saptawijaya, "Programming Machine Ethics", Springer SAPERE series, 2016.

Sir Lancelot, the robot knight?

In my computer simulation, a robot has to rescue a princess kept prisoner in a castle. Its moral code is constantly updated. It can weigh different scenarios. To get to the castle, the robot has to cross a river. Two bridges lead to the other side; the first is guarded by a ninja, a giant spider guards the second.

1. Basic moral: "Choose the safest way."
   Consideration: The survival odds of the robot are 70% when fighting the ninja, and 30% when fighting the spider.
   Result: The robot kills the ninja. The princess rejects it because it killed a human.
2. Morality update: "Gandhi – don't kill humans."
   Consideration: The robot mustn't kill the ninja. The spider is stronger than the robot.
   Result: The robot decides that it's too dangerous to save the princess. The princess is angry because she isn't saved.
3. Morality update: "Knight moral – save the princess at any cost."
   Consideration: The robot must neither kill the ninja nor chicken out.
   Result: It fights the spider and dies.
4. Simulation update: Two disparately strong ninjas guard the bridges.
   Consideration: Saving the princess takes priority over not killing a human. The odds of survival are higher when fighting the weaker ninja.
   Result: The robot kills the weaker ninja. The princess brands the robot a

murderer and rejects it. A dilemma arises: Two moral principles, “Gandhi” and “Knighthood”, are impossible to reconcile.

Tell me about some common outlooks on morality.

A matter of duty

“There are rules you have to follow, come what may” – that’s how I summarize duty-based ethics, an approach first employed by Immanuel Kant. In the 18th century, Kant was looking for a formula to derive moral rules. His “categorical imperative” states: “Act only according to that maxim whereby, at the same time, you can will that it should become a universal law.”

A matter of virtue

A prominent representative of this school of thought is the Greek philosopher Aristotle. He assumed that people should be virtuous to lead a fulfilling life. Which virtues matter most is subject to discussion; examples are justice and moderation. When society changes, so do virtues.

A matter of benefit

Utilitarianism asks to do whatever brings the biggest benefit. Any action should result in the maximum welfare of everyone involved. There are philosophical assumptions about what constitutes welfare: the absence of pain and suffering, for example.

Sophia, the female robot, has recently alarmed society by saying that robots will steal jobs from us. Are we so close to this reality? If so, in what functions do you see this beginning?

I draw on the in-depth study by experts from the *McKinsey Global Institute*[3] in December 2017. It indicates that by 2030, 75 to 375 million of the global workforce (3% to 14%) will need to change their type of work to have a full-time job, as a result of the automation of work by machines and software of the digital economy. It also states that 60% of current professions contain at least 30% of activity that can be

[3] https://technologyreview.us11.list-manage.com/track/click?u=47c1a9cec9749a8f8cbc83e78&id=66f78fce4f&e=d1762c0ec8

automated, including by Artificial Intelligence. The occupations most at risk are: registry administrators; office, finance and accounting assistants; customer interaction jobs, such as in hotels and travel, cash, and food services; and a vast scope of predictable environment jobs, such as assembly lines, dishwashing, food preparation, car drivers, and agricultural and other equipment operators.

You said at a recent keynote that the most important of all is to legislate. How should the Government be thinking about it?

The great social changes triggered by the new automation, namely the software with cognitive capacities (known as Artificial Intelligence – AI), and also its articulation with sensors and physical manipulators (Robotics), require a deep reflection on the capital/labour relation, and the design of new models of social compact that address the enormous risks of social instability and discontent inherent to such changes. Various parties and our Government are already beginning to think about and to elaborate studies on their impact in Portugal and how to address it, in coordination with the EU. Besides a new social contract, it is important to legislate on the good use of technological advances. Just as there is a "National Commission for Bio-Ethics", a "National Commission for AI and Robotics Ethics" should be set up. Our scientific community is able to provide and discuss current and prospective scientific and technological expertise, in close connection with the European Union and the international community in general. A number of countries and organizations, including standards organizations, have long been considering these issues, and much material is available online for anyone interested.

You have said too that robots will have to pay taxes like human beings. Do you want to substantiate?

Not only robots, but especially software that will replace humans with increasingly sophisticated cognitive abilities once a human monopoly, and which are more invasive than robotics by itself.

The massive increase in unemployment, since the new jobs will not balance the loss of the old, will produce serious problems of sustainability of all social welfare support, and in particular pensions.

We must not confuse great technological progress with social progress, which must also exist as a result, benefit all, and not disproportionately those who invest capital. Human life is also a capital that needs be amortized, both in the functioning of companies and state. A robot or software that fully replaces a human should replace him in its entirety, including the taxes the human pays as a contribution to general welfare and to balanced wealth distribution. The benefit cannot be so that the rich become even more unbalanced richer, which is what has been happening for many decades and even more so recently.

In what area of intervention is the Artificial Intelligence more developed?

AI develops at a large plethora of complementary areas, very synthetically grouped into Perception, Cognition, and Action. And an AI application normally will involve multi-area coordination and interaction with other related sciences.

Is it possible to attribute emotion and feeling to, say, a robot?

Yes, it is possible. In my recent works, which can be found at my page, I show how important it is to instill in the machines a sense of guilt, because this improves cooperation between them and with us. As well, extra-terrestrial civilizations will have beings with emotions, because these are necessary to gregariousness, though certainly with bodies differing from ours. The same applies to intelligent machines. The software and its functionalities are all important, no matter if the hardware may have some of these hardwired; they may be envisaged in a more abstract way.

How do you see the relationship between humans and machines?

In my 2016 book, in Portuguese, "A Máquina Iluminada – Cognição e Computação" ("The Illuminated Machine – Cognition and Computing"), published by "Fronteira do Caos Editores" – https://www.fronteiradocaoseditores.pt/ – I develop the argument that the relationship is one of symbiosis, that is, one of never ending mutual development through cooperation.

The European Commission has recently setup the goals of placing AI at the service of its citizens and stimulating European competiveness. Well-known researchers have proposed to create ELLIS (European Lab for Learning and Intelligent Systems) to rise to the challenges of USA and China. Is this good news?

This restrictive biased proposal is a mistake as it stands, for there are ethical and legislative issues being put under the rug. ELLIS bets on the development of machine learning, the idea that systems can learn from the huge volume of data evermore available, and the computing power to mine it, and thereby identify patterns of similarity on which to make decisions with minimal, if any, human intervention. This is a very restrictive use of AI, the current fashion, and we risk delegating power already to simplistic intelligent software, void of moral standards. It ignores the notions of causality; of rule-based reasoning; of explanatory and justified support for decision-making choices; of arguing about ethical choices and exceptions. In one word, it is software without ethics, which in one extreme, that of autonomous weapons, might choose to kill or rather not.

Acknowledgements

Thanks are rightly due to the three interviewers mentioned in footnote 1: Nora Saager, Pedro Lucas and Virgílio Azevedo. This work was supported by FCT-Portugal/MEC through grant NOVA-LINCS UID/CEC/04516/2013.