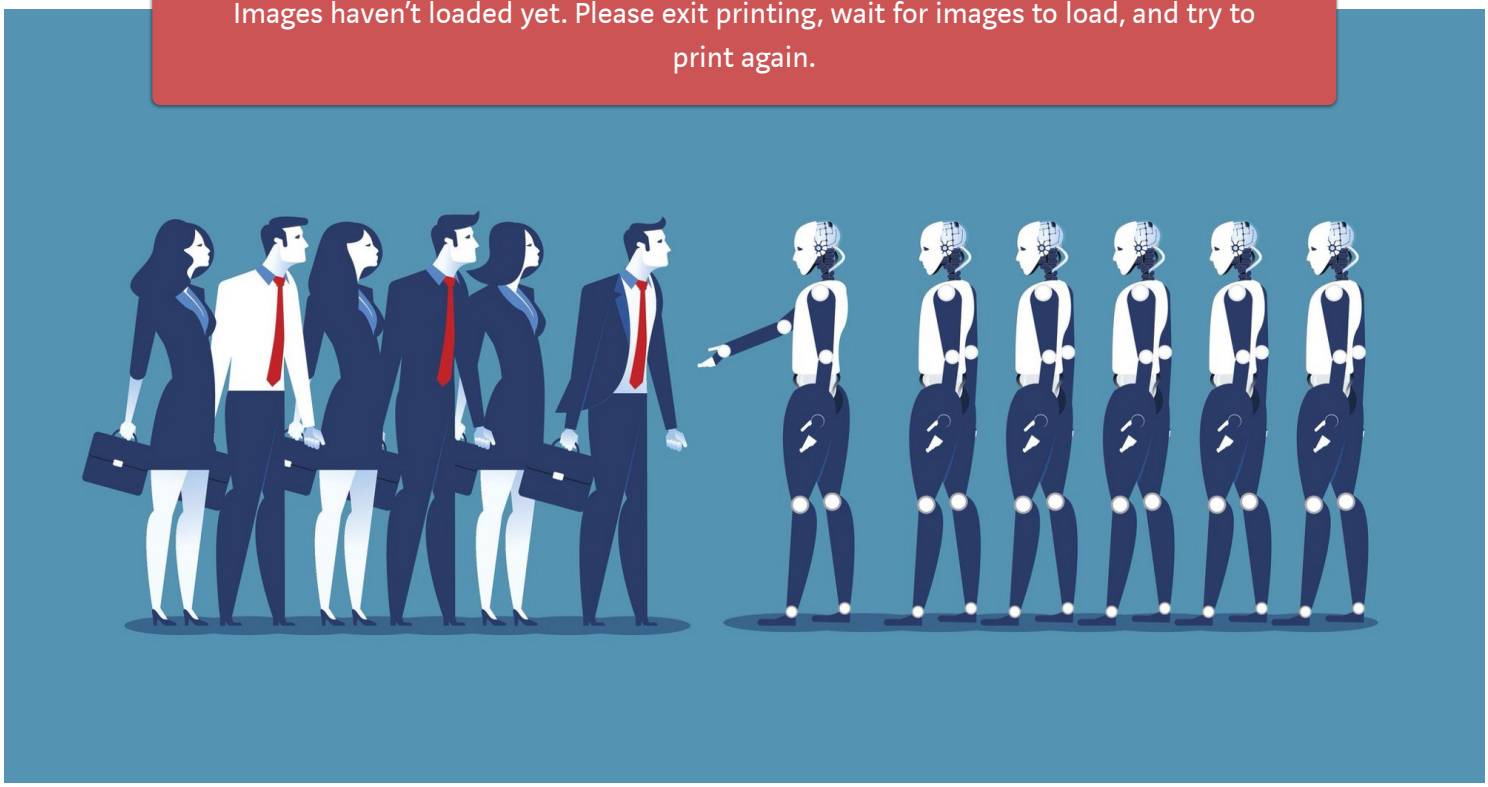


Can we teach morality to machines? Three perspectives on ethics for artificial intelligence

Before giving machines a sense of morality, humans have to first define morality in a way computers can process. A difficult but not impossible task.

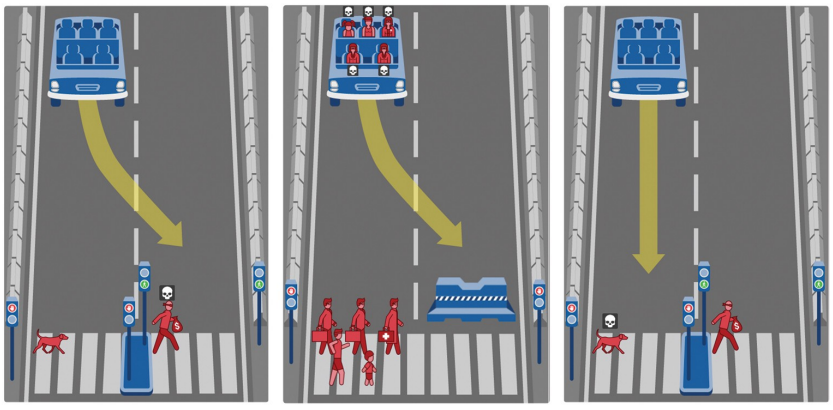
Vyacheslav Polonski, PhD Follow
Dec 19, 2017 · 7 min read ★

Today, it is difficult to imagine a technology that is as enthralling and terrifying as machine learning. While media coverage and research papers consistently tout the potential of machine learning to become the biggest driver of positive change in business and society, the lingering question on everyone's mind is: "Well, what if it all goes terribly wrong?"

For years, experts have warned against the unanticipated effects of general artificial intelligence (AI) on society. Ray Kurzweil predicts that by 2029 intelligent machines will be able to outsmart human beings. Stephen Hawking argues that "once humans develop full AI, it will take off on its own and redesign itself at an ever-increasing rate". Elon Musk warns that AI may constitute a "fundamental risk to the existence of human civilization". Alarmist views on the terrifying potential of general AI abound in the media.

More often than not, these dystopian prophecies have been met with calls for a more ethical implementation of AI systems; that somehow engineers should imbue autonomous systems with a sense of ethics. According to some AI experts, we can teach our future robot overlords to tell right from wrong, akin to a "Good Samaritan AI" that will always act justly on its own and help humans in distress.

Although this future is still decades away, today there is much uncertainty as to how, if at all, we will reach this level of general machine intelligence. But what is more crucial, at the moment, is that even the narrow AI applications that exist today require our urgent attention in the ways in which they are making moral decisions in practical day-to-day situations. For example, this is relevant when algorithms make decisions about who gets access to loans or when self-driving cars have to calculate the value of a human life in hazardous traffic situations.



Moral dilemmas for self-driving cars (Source: MIT Media Lab)

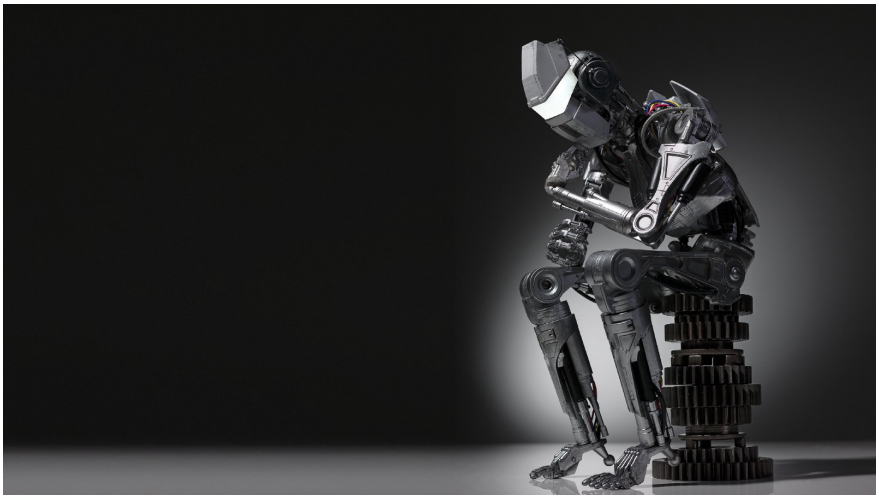
Moral problems in everyday life

Teaching morality to machines is hard because humans can't objectively convey morality in measurable metrics that make it easy for a computer to process. In fact, it is even questionable whether we, as humans have a sound understanding of morality at all that we can all agree on. In moral dilemmas, humans tend to rely on gut feeling instead of elaborate cost-benefit calculations. Machines, on the other hand, need explicit and objective metrics that can be clearly measured and optimized.

For example, an AI player can excel in games with clear rules and boundaries by learning how to optimize the score through repeated playthroughs. After its experiments with deep reinforcement learning on Atari video games, Alphabet's DeepMind was able to beat the best human players of Go. Meanwhile, OpenAI amassed "lifetimes" of experiences to beat the best human players at the Valve Dota 2 tournament, one of the most popular e-sports competitions globally.

But in real-life situations, optimization problems are vastly more complex. For example, how do you teach a machine to algorithmically maximise fairness or to overcome racial and gender biases in its training data? A machine cannot be taught what is fair unless the engineers designing the AI system have a precise conception of what fairness is.

This has led some authors to worry that a naive application of algorithms to everyday problems could amplify structural discrimination and reproduce biases in the data they are based on. In the worst case, algorithms could deny services to minorities, impede people's employment opportunities or get the wrong political candidate elected. Some people have argued that the use of AI in politics already had disastrous consequences.



Thinking about new ways to teach robots right from wrong.

So what can we do about it? Based on our experiences in machine learning, we believe there are three ways to begin designing more ethically aligned machines with the following guidelines:

1. Explicitly defining ethical behaviour

AI researchers and ethicists need to formulate ethical values as quantifiable parameters. In other words, they need to provide machines with explicit answers and decision rules to any potential ethical dilemmas it might encounter. This would require that humans agree among themselves on the most ethical course of action in any given situation—a challenging but not impossible task. For example, Germany's Ethics Commission on Automated and Connected Driving has recommended to specifically programme ethical values into self-driving cars to prioritize the protection of human life above all else. In the event of an unavoidable accident, the car should be "prohibited to offset victims against one another". In other words, a car shouldn't be able to choose whether to kill one person based on individual features, such as age, gender or physical/mental constitution when a crash is inescapable.

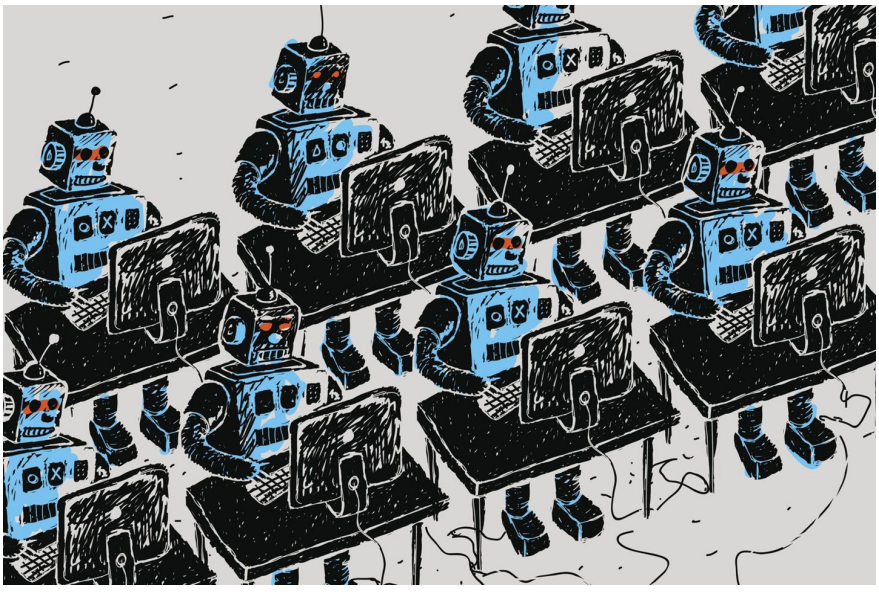
2. Crowdsourcing human morality

Engineers need to collect enough data on explicit ethical measures to appropriately train AI algorithms. Even after we have defined specific metrics for our ethical values, an AI system might still struggle to pick it up if there is not enough unbiased data to train the models. Getting appropriate data is challenging, because ethical norms cannot be always clearly standardized. Different situations require different

ethical approaches, and in some situations there may not be a single ethical course of action at all—just think about lethal autonomous weapons that are currently being developed for military applications. One way of solving this would be to crowdsource potential solutions to moral dilemmas from millions of humans. For instance, MIT's Moral Machine project shows how crowdsourced data can be used to effectively train machines to make better moral decisions in the context of self-driving cars.

3. Making AI systems more transparent

Policymakers need to implement guidelines that make AI decisions with respect to ethics more transparent, especially with regard to ethical metrics and outcomes. If AI systems make mistakes or have undesired consequences, we cannot accept "the algorithm did it" as an adequate excuse. But we also know that demanding full algorithmic transparency is technically untenable (and, quite frankly, not very useful). Neural networks are simply too complex to be scrutinized by human inspectors. Instead, there should be more transparency on how engineers quantified ethical values before programming them, as well as the outcomes that the AI has produced as a result of these choices. For self-driving cars, for instance, this could imply that detailed logs of all automated decisions are kept at all times to ensure their ethical accountability.



How can moral values be measured and optimised?

Next steps for moral machines

We believe that these three recommendations should be seen as a starting point for developing ethically aligned AI systems. Failing to imbue ethics into AI systems, we may be placing ourselves in the dangerous situation of allowing algorithms to decide what's best for us. For example, in an unavoidable accident situation, self-driving cars will need to make some decision for better or worse. But if the car's designers fail to specify a set of ethical values that could act as decision guides, the AI system may come up with a solution that causes more harm.

This means that we cannot simply refuse to quantify our values. By walking away from this critical ethical discussion, we are making an implicit moral choice. And as machine intelligence becomes increasingly pervasive in society, the price of inaction could be enormous—it could negatively affect the lives of billions of people.

Machines cannot be assumed to be inherently capable of behaving morally. Humans must teach them what morality is, how it can be measured and optimised. For AI engineers, this may seem like a daunting task. After all, defining moral values is a challenge mankind has struggled with throughout its history. If we can't agree on what makes a moral human, how can we design moral robots?

Nevertheless, the state of AI research and its applications in society require us to finally define morality and to quantify it in explicit terms. This is a difficult but not impossible task. Engineers cannot build a "Good Samaritan AI", as long as they lack a formula for the Good Samaritan human.

. . .

About the authors: Jane Zavalishina is the CEO of Yandex Data Factory, a provider of AI-based solutions for industrial companies. Jane is a frequent speaker on the topics of AI business strategy and applications at various events in Europe, Middle East and Asia. She serves on the World Economic Forum's Global Future Councils. In 2016, Jane was named in Silicon Republic's Top 40 Women in Tech as an Inspiring Leader and recognised by Inspiring Fifty as one of the top 50 most inspirational women in the technology sector in the Netherlands.

Dr Vyacheslav Polonski is a researcher at the University of Oxford, studying complex social networks and collective behaviour. He holds a PhD in computational social science and has previously studied at Harvard, Oxford and LSE. He is the founder and CEO of Avantgarde Analytics, a machine learning startup that harnesses AI and behavioural psychology for the next generation of algorithmic campaigns. Vyacheslav is actively involved in the World Economic Forum Expert Network and the WEF Global Shapers community, where he served as the Curator of the Oxford Hub. He writes about the intersection of sociology, network science and technology.

Earlier versions of this article were published on the Net Politics Blog of the Council on Foreign Relations on 14 November 2017, the World Economic Forum Agenda on 23 November 2017 and the official blog of the BCG Centre for Public Impact on 12 December 2017. The article was also translated into French and Polish in other online media outlets.