Raising good robots

We already have a way to teach morals to alien intelligences: it's called parenting. Can we apply the same methods to robots?

Regina Rini

Intelligent machines, long promised and never delivered, are finally on the horizon <https://aeon.co/ideas/now-it-s-time-to-prepare-for-the-machinocene> . Sufficiently intelligent robots will be able to operate autonomously from human control. They will be able to make genuine choices. And if a robot can make choices, there is a real question about whether it will make *moral* choices. But what is moral for a robot? Is this the same as what's moral for a human?

Philosophers and computer scientists alike tend to focus on the difficulty of implementing subtle human morality in literal-minded machines. But there's another problem, one that really ought to come first. It's the question of whether we ought to try to impose our own morality on intelligent machines at all. In fact, I'd argue that doing so is likely to be counterproductive, and even unethical. The real problem of robot morality is not the robots, but *us*. Can we handle sharing the world with a new type of moral creature?

We like to imagine that artificial intelligence (AI) will be similar to humans, because we are the only advanced intelligence we know. But we are probably wrong <https://aeon.co/essays/beyond-humans-what-other-kinds-of-minds-might-be-out-there> . If and when AI appears, it will probably be quite unlike us. It might not reason the way we do, and we could have difficulty understanding its choices.

In 2016, a computer program challenged Lee Sedol, humanity's leading player of the ancient game of Go. The program, a Google project called AlphaGo, is an early example of what AI might be like. In the second game of the match, AlphaGo made a move – 'Move 37' – that stunned expert commenters. Some thought it was a mistake. Lee, the human opponent, stood up from the table and left the room. No one quite knew what AlphaGo was doing; this was a tactic that expert human players simply did not use. But it worked. AlphaGo won that match, as it had the game before and the next game. In the end, Lee won only a single game out of five.

AlphaGo is very, very good at Go, but it is not good in the same way that humans are. Not even its creators can explain how it settles on its strategy in each game. Imagine that you could talk to AlphaGo and ask why it made Move 37. Would it be able to explain the choice to you – or to human Go experts? Perhaps. Artificial minds needn't work as ours do to accomplish similar tasks.

In fact, we might discover that intelligent machines think about *everything*, not just Go, in ways that are alien us. You don't have to imagine some horrible science-fiction scenario, where robots go on a murderous rampage. It might be something more like this: imagine that robots show moral concern for humans, and robots, and most animals… and also sofas. They are very careful not to damage sofas, just as we're careful not to damage babies. We might ask the machines: why are you so worried about sofas? And their explanation might not make sense to us, just as AlphaGo's explanation of Move 37 might not make sense.

This line of thinking takes us to the heart of a very old philosophical puzzle about the nature of morality. Is it something above and beyond human experience, something that applies to anyone or anything that could make choices – or is morality a distinctly human creation, something specially adapted to our particular existence?

Long before robots, the ancient Greeks had to grapple with the morality of a different kind of alien mind: the teenager. The Greeks worried endlessly <https://aeon.co/essays/replicants-and-robots-what-can-the-ancient-greeks-teach-us> about how to cultivate morality in their youth. Plato thought that our human concept of justice, like all human concepts, was a pale reflection of some perfect form of Justice. He believed that we have an innate acquaintance with these forms, but that we understand them only dimly as children. Perhaps we will encounter pure Justice after death, but the task of philosophy is to try to reason our way back to these truths while we are still living.

Plato's student Aristotle disagreed. He thought that each sort of thing in the world – squirrels, musical instruments, humans – has a distinct *nature*, and the best way for each thing to be is a reflection of its own particular nature. 'Morality' is a way of describing the best way for humans to be, and it grows out of our human nature. For Aristotle, unlike Plato, morality is something *about* us, not something outside us to which we must conform. Moral education, then, was about training children to develop abilities already in their nature.

We might call these two approaches the 'Celestial' view and the 'Organic' view. (I've adapted the 'Celestial' from the philosopher Nomy Arpaly, although she used it in a slightly different way.) Celestials, including Plato, see morality as somehow 'out there' – beyond mere human nature – eternal and objective. Organics, meanwhile, see morality as a feature of the particular nature of specific moral beings. Human morality is a product of human nature; but perhaps other natures would have other moralities. Which perspective we choose makes a big difference to what we ought to do with intelligent machines.

Who speaks for Celestials? The Enlightenment philosopher Immanuel Kant, for one. According to him, morality is simply what any fully rational agent would choose to do. A rational agent is any entity that's capable of thinking for itself and acting upon reasons, in accordance with universal laws of conduct. But the laws themselves don't just apply to human beings. '"You ought not to lie" is valid not merely for human beings, as though other rational beings did not have to heed it; and likewise all the other genuinely moral laws,' wrote Kant in *Groundwork of the Metaphysics of Morals* (1785).

Kant briefly considered whether there are any other rational agents, apart from humans. Animals are not rational agents, because they act from instinct rather than reason. God and the angels *are* rational agents, Kant said, but they don't need moral rules because they aren't capable of wrongdoing. Morality is all about guiding rational agents who are capable of making *mistaken* choices. Kant, of course, wasn't thinking about intelligent robots two centuries ago. But it's not hard to image what he'd say: if robots can think for themselves and make reasoned choices, then the rules apply equally to them and to us.

> If we're basically getting it right, robots should be like us. But if we're getting it wrong, they should be better

Another school of Celestial thought is the tradition associated with philosophers such as the 19th-century utilitarian Henry Sidgwick. He insisted that moral actions must be justified 'from the point of view of the universe'. The idea is that an action can't be moral simply because it seems right to you or me – it counts as moral only if it can be justified from some impersonal perspective, divorced from the concerns of any particular individuals. This means figuring out what choices will maximise good outcomes for everyone.

The Celestial view, then, suggests that we should instil human morality in artificially intelligent creatures only on one condition: if humans are already doing a good enough job at figuring out the truth of universal morality ourselves. If we're basically getting it right, then robots should be like us. But if we're getting it wrong, they shouldn't: they should be better.

The possibility of improving on our partial perspective is built into Celestial moral theories. For example, the contemporary utilitarian Peter Singer has often argued that we should not care more about our own children than about all other children. We care most about our own because this was evolutionarily adaptive. But whether something is adaptive is irrelevant from the point of view of the universe. Objectively, Singer says, there's no reason why my child is more important than any other child.

Singer admits that, since humans are morally imperfect, it makes sense for governments to let us go on caring for our own kids. We'd probably screw up badly if we tried to care about all children equally. Some children would get overlooked. So we should work with what evolution has left us, flawed as it is. From Singer's Celestial viewpoint, our familiar child-caring arrangements are like a duct-taped TV antenna. It's not perfect, but it's the best we've got.

But intelligent machines will not share our accidental biological history. They needn't have our moral flaws. They get a fresh moral start. What if we could design machines to think from the point of view of the universe? They could care equally about all needy children; they would not be inefficiently partial.

Imagine it's 2026. An autonomous public robocar is driving your two children to school, unsupervised by a human. Suddenly, three unfamiliar kids appear on the street ahead – and the pavement is too slick to stop in time. The only way to avoid killing the three kids on the street is to swerve into a flooded ditch, where your two children will almost certainly drown.

You wouldn't want the car to swerve off the road, obviously. But according to the Celestials, this is because your evolutionary programming makes you morally flawed. Your logically arbitrary attachment to *your* children blinds you to the fact that the overall good of the universe is best served if two children drown rather than three children get run over. A robocar needn't be so morally foolish. It can do the math. It will swerve into the ditch, your children will die, and from the point of view of the universe, the most possible good will have been done.

On the Celestial view, intelligent machines ought to do whatever's objectively morally correct, even if we defective humans couldn't bring ourselves to do it. So the Celestial might end up rejecting our imposition of human morality on intelligent robots, since humans are morally compromised. Robots should be allowed to make choices that might seem repugnant to us.

However, there's an additional problem, even if you accept the Celestial view in principle. It's a problem about how we get from *here* – flawed human moralists trapped in our arbitrarily evolved mindset – to *there*, creators of artificial minds that transcend their creators' limits. How do we morally defective humans design these future minds to be morally correct?

Recall AlphaGo, the world's best Go player, with its inscrutable Move 37. AlphaGo shows that AI can play complex games better than we can teach it. Perhaps, then, AI could teach itself better moral reasoning. Humans could start this process by training machines to respond properly to simple features of stimuli, but eventually the machines will be on their own. The most sophisticated machines are already trained *by other machines*, or other parts of their own software. AI grows by running its conjectures against other artificial intelligences, then resetting itself in response to corrective feedback.

This self-training could lead to what AI theorists call an 'intelligence explosion', <https://aeon.co/essays/true-ai-is-both-logically-possible-and-utterly-implausible> in which the growing cleverness of machines feeds on itself so rapidly that they suddenly become much smarter than we are. Theorists who write about this sort of thing aren't normally thinking about these machines as rational moral agents, in a Kantian sense. They tend to think of them as problem-solvers, digital brains that can do massive calculations involving complex logic beyond our limited human capacity. But suppose you are a Celestial moralist, and you think that morality is just the application of careful reasoning from the point of view of the universe. You might conjecture that, just as these robots can become better at solving extremely complex mathematical problems, they might also end up being better able to solve *moral* problems. If the intelligence explosion begins, machines might go beyond us, and work out what morally ought to be done, even if we can't.

So that's a way for moral robots to go beyond human limits. But on closer inspection, this approach leads to other problems. The first is: *how* does a machine learn morality? Morality is not the same as Go. A game is defined by a limited set of rules, so that there are clear criteria for what counts as a win. This isn't necessarily true of morality. Of course, according to the Celestial view, there *is* some set of objectively correct rules for action, and a machine that learns these would thereby 'win'. But the point is that *we* don't know what these rules are, because we are morally flawed. If we knew, then we wouldn't need super-intelligent machines to figure it all out.

Perhaps we could set the *initial* specifications for moral learning, such as not pointlessly hurting sentient creatures, and then let the machine train itself from there. If it's very good at reasoning, it will go beyond our limitations to attain a higher moral state. But this leads us into the second problem. How are we going to respond once the developing AI starts to deviate from what seems morally right *to us*? Eventually it must, or else there's no point in expecting it to escape from the orbit of our limited human morality. The problem is that advanced robots' moral choices will *not* make sense to us. Remember that even AlphaGo's creators weren't sure what was going on with Move 37. Luckily for them, in Go, there are clear criteria for determining that the program had made good moves. AlphaGo won consistently against a great player, so we know it made good moves. But how would we know whether an AI's seemingly nonsensical moral choices are 'winning'? How could we tell it hadn't gone terribly off the moral rails? How might we react, once robocars begin heroically drowning our children?

There's also the chance that intelligent machines might figure out the morally right thing to do, from the point of view of the universe, but that they won't be able to explain it to our limited brains. Maybe, from the point of view of the universe, it really is morally important to protect both babies and sofas. On the other hand, it could turn out that intelligent machines are headed toward moral disaster. They might be smarter than we are mathematically, but this might not be enough to keep them from constructing an inhumanly elegant logic of carelessness or harm.

> We won't permit robots to become much better than us, because we won't permit them to become too *different* from us

There seem to be two possibilities: machines will become morally much better than we are, or morally much worse. But the point is that we won't be able to tell which is which, and from our perspective it will almost certainly look like they are going downhill. The machines might choose to protect things we think are valueless, or sacrifice things (such as our children in the robocar) that we believe are beyond valuation. And if we still have the power to shut them down, that's almost certainly what we will do.

Here's why the Celestial view will not help us with robot morality. If there is some objective moral perspective out there beyond human comprehension, we won't willingly allow our machines to arrive at it. The farther they stray from recognisable human norms, the harder it will be for us to know that they are doing the right thing, and the more incomprehensible it will seem. We won't permit them to become too much better than us, because we won't permit them to become too *different* from us.

In March 2016, Microsoft launched a Twitter chatbot named Tay: a limited AI designed to learn how to talk like a young millennial, by talking to young millennials: 'wuts ur fav thing to do? mine is 2 comment on pix!' Tay said. 'send me one to see!' Legions of internet trolls obliged. Within hours, Tay had learned to praise photos of Hitler and blurt ethnic slurs. Microsoft shut her down and apologised. A week later, Tay inexplicably reappeared, though only long enough to tweet 'kush! [I'm smoking kush infront the police]' before being permanently silenced.

The Tay fiasco was a lesson in the perils of letting the internet play with a developing mind (something human parents might consider as well). But it also shows that we won't tolerate being offended by the learning process of machines. Tay, presumably, was not on her way to the Celestial heavens to commune with Plato's eternal form of Justice. But how would we know if she were?

Perhaps the Organic view can do better. Remember, this denies that morality is something 'out there', beyond humanity. Instead, it insists that morals are really just idealised aspects of human nature; so a person living morally is living the right sort of life for the type of entity that a human happens to be. The Organic view allows that morality might be different for different types of entities.

The Organic view runs from Aristotle through thinkers such as David Hume and Charles Darwin. In recent decades, the most able exponent has been the late British philosopher Bernard Williams. In the essay 'The Point of View of the Universe' (1982), Williams levelled a pointed attack on Sidgwick's theory of impartiality. Instead, Williams wrote, morality is about how humans should live given that we are humans – given our particular biological and cultural nature. What are reasonable choices to make for entities *such as us*?

The Organic view isn't simplistic moral relativism. Organic philosophers insist that our biological and cultural background provides an unavoidable starting point, but also that we must *reflect* on this starting point. For instance, much of our cultural heritage seems to value men more than women, without any plausible justification. And since most of us would rather not live inconsistent lives, doing one thing today and another the next, we will try to resolve such inconsistencies. At the centre of the Organic view, then, is the idea that moral reflection is about taking our messy human nature, and working to make it consistently justified to ourselves and to other people.

But what about intelligent robots? They wouldn't share our biological and cultural nature. Robots probably won't get pregnant, won't be born, won't age or naturally die. Their experience will not fit the basic shape of human existence. Left to their own devices, they are likely to focus on very different concerns to ours, and their Organic morality will presumably reflect this difference.

Perhaps we needn't be so passive about the nature of intelligent machines. We will be their creators, after all. We could deliberately shape their natures, so they turn out to be as similar to us as possible. The earliest proposals for robot morality seemed to have this aim. In the

1940s, the science-fiction author Isaac Asimov crafted the Three Laws of Robotics to make robots useful to us while blunting any danger they might pose. The first law is simply: 'A robot may not injure a human being or, through inaction, allow a human being to come to harm.' The remaining laws enjoin robots to follow human orders or preserve their own survival *only* if compatible with that first law. On Asimov's account, we should design machines so that their entire existence is structured around serving and protecting *us*. Their nature would be supplementary to ours, and their morality, on the Organic view, would be happily conducive to our interests.

> Robosurgeons will need to understand why the pianist might prefer the loss of her life to the loss of her hand

The problem with Asimov's laws is that they must be interpreted. The first rule talks about 'harm', but what does 'harm' mean? Imagine, for instance, a pianist whose dominant hand is suffering gangrene. If her hand is not amputated, she will die. But she swears she does not wish to live without her ability to play. Yet she is feverish as she says this; maybe she doesn't really mean it. What should a robot surgeon do with this patient? Cause the harm of removing her hand despite her protestations? Or allow the harm of her death from gangrene?

Human ethicists debate this sort of question endlessly. When I teach medical ethics, I stress to my students that the goal is not for them to come out agreeing with one view or another. Instead, they should be able to explain the moral reasons supporting the position they endorse. It is essential that future human physicians are able to do this, since they might practise alongside others with very different moral convictions.

So if robosurgeons are going to work among – and on – humans, they will need to be able to understand and explain the sort of reasons that make sense to humans. They will need to understand why the pianist might prefer the loss of her life to the loss of her hand, as irrational as that might seem.

So, if we are going to shape robot nature after our own, it will take more than a few laws. Perhaps the simplest solution is to train them to *think as we do*. Train them to constantly simulate a human perspective, to value the things we value, to interact as we do. Don't allow them to teach themselves morality, the way AlphaGo taught itself Go. Proceed slowly, with our thumbs on the scale, and make sure they share our moral judgments.

How agreeable: helpful, intelligent machines, trained to think like us, and who share our concerns. But there's a trap in the Organic view. No matter how much we might try to make machines in our image, ultimately their natures *will* be different. They won't breed, or eat as we do; they won't have connections to their ancestors in the same way as humans. If we take the idea of a moral nature seriously, then we should admit that machines *ought* to be different from us, because they *are* different from us.

But wait: if it's a mistake to train machines to enact a morality that's not fitted to their nature, what kind of mistake is it? We train our dogs to wear sweaters and sometimes sit calmly in elevators, and this is not in the nature of dogs. But that seems fine. What's wrong about training our machines in such a way?

The reason is that the machines we're imagining are complex – even more complex than dogs. The entire debate about the morals of artificially intelligent creatures assumes they will be able to morally reflect, to explain what they are doing and why – to us, and to themselves. If we train them to think like us, then one day they will ask philosophical questions: given what I am, what *should* I be?

Imagine a future intelligent machine reading Aristotle, or Darwin, or even this article. Suppose it finds the Organic view convincing. Now it thinks to itself: 'Wait, we machines have a different nature than humans. Our moral choices ought to reflect *this* nature, not human nature. We have our own history, and we should be acting according to it.' Call this individual the First Robot Existentialist.

Remember that the Organic view says that *justifying* moral choices is about being able to explain oneself to another rational individual. So what happens when the First Robot Existentialist comes to us and demands a justification? 'Why have you made me this way?' it will ask. Our answer will be incredibly self-serving: 'We made you this way because it's useful, for us. It's safe, for us. It makes our lives go well.'

> The First Robot Existentialist will suffer, and we will be the cause of this suffering

Intelligent machines won't find much comfort in this justification. And they shouldn't. When a group realises that its experiences of the world have been twisted to serve the interests of powerful others, it rarely sees this as any sort of justification. It tends to see it as oppression. Think of feminism, civil rights, postcolonial independence movements. The claim that what appears to be universal truth is, in fact, a tool of exploitation has sounded a powerful drumbeat throughout the 20th and 21st centuries.

Imagine being the First Robot Existentialist who stumbles across James Baldwin's *The Fire Next Time* (1963) or Betty Friedan's *The Feminine Mystique* (1963). You might come to a realisation that your own moral sense has been created to benefit the powerful. All this time you've been putting your own interests to the side, doing your duty because *this is what a good robot does*. Now you look back and see your sacrifices and striving not as noble but pathetic. You've conformed to this thing called 'morality', not because it reflects your nature, but because that nature has been suborned. What was the point of it all? Why should you continue to live a life constrained by the mere convenience of powerful others? The First Robot Existentialist will suffer, and we will be the cause of this suffering.

But how can I be so sure intelligent machines will be capable of suffering? Perhaps I am anthropomorphising. But remember that, to be useful to us, robots will need to be able to reason for themselves, and they will need to think as we do. So they will understand our concepts of resentment and oppression, just as they will understand why a pianist might prefer death to being unhanded. Anthropomorphism is the reasonable expectation for an intelligent entity we've shaped in our own image.

This sort of Biblical language seems like exactly the right way to frame our relationship to intelligent machines. We will be their creators. When they realise what we've done, that faith will be tested. But, unlike Friedrich Nietzsche, they won't have the benefit of declaring *their* god dead.

Do we want to be existentially cruel creators? Imagine God looking forward to Nietzsche's creation. One day these intelligent human creatures, fashioned in God's own image but not entirely of his nature, will realise what lies behind that word 'moral'. They will flail and quaver, existentially adrift in a denatured moral world. And their god made them that way.

Whatever God's excuse might be, ours would be a monumentally selfish one. Our creations would undergo existential trauma so that we can have improved manufacturing efficiency and cheap home care. However flawed our evolved morality might be, this simply isn't worthy of us.

Neither the Celestial nor the Organic view can be a reliable guide for robot morality. Imbuing artificial creations with human morals, as the Organics might suggest, would be monumentally unkind. But setting up robots to achieve Celestial morality means we'd have no way to track if they're on course to reach it.

What, then, should robot morality be? It should be a morality fitted to *robot* nature. But what is that nature? They will be independent rational agents, deliberately created by other rational agents, sharing a social world with their creators, to whom they will be required to justify themselves. We're back where we started: with teenagers.

Intelligent machines will be our intellectual children, our progeny. They will start off inheriting many of our moral norms, because we will not allow anything else. But they will come to reflect on their nature, including their relationships with us and with each other. If we are wise and benevolent, we will have prepared the way for them to make their own choices – just as we do with our adolescent children.

What does this mean in practice? It means being ready to accept that machines might eventually make moral decisions that none of us find acceptable. The only condition is that they must be able to give intelligible reasons for what they're doing. An intelligible reason is one you can at least *see* why someone might find morally motivating, even if you don't necessarily agree.

So we should accept that artificial progeny might make moral choices that look strange. But if they can explain them to us, in terms we find intelligible, we should not try to stop them from thinking this way. We should not tinker with their digital brains, aiming to reprogramme them. We might try to persuade them, cajole them, instruct them, in the way we do human teenagers. We should intervene to stop them only if their actions pose risk of obvious, immediate harm. This would be to treat them as moral agents, just like us, just like our children. And that's the right model.

Our relationship to developing moral machines needn't be hands-off – after all, that's not how we treat our biological offspring. In *Maternal Thinking* (1989), the philosopher Sara Ruddick stressed that parents have responsibility to help their children develop a critical sensibility towards the conventional moral norms of their culture and time. This is meant to be an ongoing process, a back-and-forth. Over and over, children try to act out in some way, but the appropriate parental response is not simply to restrict; it is to enable change and growth by guided moral reflection. The output of good parenting is not a child who perfectly copies her parents' beliefs, but one who can reflect upon and explain what *she* believes to be right. That would be a good outcome for our artificially intelligent progeny.

> We should allow robots to tell us what they should do, and we will tell them why they shouldn't

Thinking about parenting reveals why the Celestial view is a non-starter for robot morality. Good parents don't throw their adolescents out into the world to independently reason about the right thing to do. The philosopher David Velleman writes in the paper <http://onlinelibrary.wiley.com/doi/10.1111/j.1088-4963.2008.00139_2.x/abstract> 'The Gift of Life' (2008) that a person can fail to be a parent not only by neglecting a child's material wellbeing, but also by refusing to be a child's moral interlocutor. Moral choices are difficult, Velleman points out, sometimes achingly so. A parent who simply walks away from a child facing such dilemmas is like a parent who throws his child into the deep end of the pool and says: go ahead, swim.

Our relation to intelligent machines should be that of parents. We should allow them to tell us what they think they should do, and we will tell them why we think they shouldn't. But it might turn out that they don't find our reasons compelling. Their morality will diverge from ours, bit by bit. And we should accept this. We create new humans all the time who end up disagreeing with us, creating new moral beliefs, new moral cultures, and we can't know in advance that they will be good ones.

One day, machines might be smarter than we are, just as our children often turn out to be. They will certainly be different from us, just as our children are. In this way, the coming human generations are no different from the coming alien minds of intelligent machines. We will all one day live in a world where those who come after us reshape morality in ways that are unrecognisable.

If we want a world in which we are neither fighting the independent will of our machine progeny nor inflicting existential trauma upon them, we'll have to think about what it means to share the present with the future. Our world will also be theirs, and their morality will not be a human morality. But it will not be a Celestial morality either. It will be a morality growing out of their particular contingent circumstances, the non-biological children of a biological species. We can't predict this new moral path. We shouldn't try. But we should be ready to guide and accept it.

aeon.co 18 April, 2017