**How to prevent AI from taking over the world**

If we can’t teach machines to internalise human values and make decisions based on them, we must accept – and ensure – that AI is of limited use to us.

BY RUTH CHANG



**Sign Up**

Get the New Statesman’s Morning Call email.

[**SIGN-UP**](https://www.newstatesman.com/morning-call-email-newsletter)

Right now AI [diagnoses cancer](https://www.bbc.com/news/health-50857759#:~:text=Artificial%20intelligence%20is%20more%20accurate,images%20from%20nearly%2029%2C000%20women), decides whether you’ll [get your next job](https://www.vox.com/recode/2019/12/12/20993665/artificial-intelligence-ai-job-screen), approves your [mortgage](https://siliconangle.com/2020/10/19/googles-lending-docai-service-uses-machine-learning-fast-track-mortgage-applications/), [sentences felons](https://www.theatlantic.com/ideas/archive/2019/06/should-we-be-afraid-of-ai-in-the-criminal-justice-system/592084/), [trades](https://insidebitcoins.com/ai-trading) on the stock market, populates your [news feed](https://www.mic.com/articles/178119/how-artificial-intelligence-and-deep-learning-secretly-control-what-you-see-on-facebook), protects you from[fraud](https://cio.economictimes.indiatimes.com/news/digital-security/how-ai-ml-can-help-financial-institutions-with-fraud-prevention/71878088) and [keeps you company](https://www.bbc.com/worklife/article/20200325-can-voice-technologies-using-ai-fight-elderly-loneliness#:~:text=Voice%20technologies%20using%20AI%20are,including%20Sweden%20and%20the%20UK.&text=Loneliness%20is%20a%20global%20problem,day%20or%20being%20severely%20overweight.) when you’re feeling down.

Soon it will drive you to town, deciding along the way whether to swerve to avoid hitting a wayward fox. It will also tell you how to schedule your day, which career best suits your personality and even how many children to have.

In the further future, it could cure cancer, eliminate poverty and disease, wipe out crime, halt global warming and help colonise space. Fei-Fei Li, a leading technologist at Stanford, paints a rosy picture: “I imagine a world in which AI is going to make us work more productively, live longer and have cleaner energy.” General optimism about AI is shared by [Barack Obama](https://www.wired.com/2016/10/president-obama-mit-joi-ito-interview/), [Mark Zuckerberg](https://money.cnn.com/2017/07/25/technology/elon-musk-mark-zuckerberg-ai-artificial-intelligence/index.html) and [Jack Ma](https://www.cnbc.com/2019/08/30/jack-ma-ai-could-give-us-12-hour-work-weeks.html), among others.

And yet from the beginning AI has been dogged by huge concerns.

What if AI develops an intelligence far beyond our own? Stephen Hawking [warned](https://www.upi.com/Top_News/World-News/2017/11/07/Stephen-Hawking-warns-AI-could-develop-will-of-its-own-to-destroy-humanity/3671510038720/) that “AI could develop a will of its own, a will that is in conflict with ours and which could destroy us.” We are all familiar with the typical plotline of dystopian sci-fi movies. An alien comes to Earth, we try to control it, and it all ends very badly. AI may be the alien intelligence already in our midst.

A new algorithm-driven world could also [entrench and propagate injustice](https://en.wikipedia.org/wiki/Algorithms_of_Oppression), while we are none the wiser. This is because the algorithms we trust are often black boxes whose operation we don’t – and [sometimes *can’t*](https://www.wired.com/story/our-machines-now-have-knowledge-well-never-understand/) – understand. Amazon’s now infamous facial recognition software, Rekognition, seemed like a mostly innocuous tool for landlords and employers to run low-cost background checks. But it was [seriously biased](https://www.aclu.org/blog/privacy-technology/surveillance-technologies/amazons-face-recognition-falsely-matched-28) against people of colour, matching 28 members of the US Congress – disproportionately minorities  – with profiles stored in a database of criminals. AI could perpetuate our worst prejudices.

Finally, there is the problem of what happens if AI is too good at what it does. Its beguiling efficiency could seduce us into allowing it to make more and more of our decisions, until we “forget” how to make good decisions on our own, in much the way we rely on our smartphones to remember phone numbers and calculate tips. AI could lead us to abdicate what makes us human in the first place – our ability to take charge of and tell the stories of our own lives.

***[See also:***[***Philip Ball on how machines think***](http://www.newstatesman.com/science-tech/technology/2019/12/how-do-machines-think)***]***

It’s too early to say how our technological future will unfold. But technology heavyweights such as [Elon Musk](https://www.theverge.com/2020/2/18/21142489/elon-musk-ai-regulation-tweets-open-ai-tesla-spacex-twitter" \t "_blank) and [Bill Gates](https://www.youtube.com/watch?v=C3ZbTDA1YiQ) agree that we need to do something to control the development and spread of AI, and that we need to do it now.

Obvious hacks won’t do. You might think that we can control AI by pulling its plug. But experts warn that a super-intelligent AI could easily predict our feeble attempts to shackle it and undertake measures to protect itself by, say, storing up energy reserves and infiltrating power sources. Nor will encoding a master command, “Don’t harm humans”, save us, because it’s unclear what  “harm” means or constitutes. When your self-driving vehicle swerves to avoid hitting a fox, it exposes you to a slight risk of death – does it thereby harm you? What about when it swerves into a small group of people to avoid colliding with a larger crowd?

\*\*\*

The best and most direct way to control AI is to ensure that its values are *our*values. By building human values into AI, we ensure that everything an AI does meets with our approval. But this is not simple. The so-called “[Value Alignment Problem](https://brianchristian.org/the-alignment-problem/)” – how to get AI to respect and conform to human values – is arguably the most important, if vexing, problem faced by AI developers today.

So far, this problem has been seen as one ofuncertainty: if only we understood our values better, we could program AI to promote these values. [Stuart Russell](http://people.eecs.berkeley.edu/~russell/), a leading AI scientist at Berkeley, offers an intriguing solution. Let’s design AI so that its goals are unclear. We then allow it to fill in the gaps by observing human behaviour. By learning its values from humans, the AI’s goals will be our goals.

This is an ingenious hack. But the problem of value alignment isn’t an issue of technological design to be solved by computer scientists and engineers. It’s a problem of human understanding to be solved by philosophers and axiologists.

The difficulty isn’t that we don’t know enough about our values – though, of course, we don’t. It’s that even if we had full knowledge of our values, these values might not be computationally amenable*.*If our values can’t be captured by algorithmic architecture, even approximately, then even an omniscient God couldn’t build AI that is faithful to our values. The basic problem of value alignment, then, is what looks to be a fundamental mismatch between human values and the tools currently used to design AI.

Paradigmatic AI treats values as if they were quantities like length or weight – things that can be represented by cardinal units such as inches, grams, dollars. But the pleasure you get from playing with your puppy can’t be put on the same scale of cardinal units as the joy you get from holding your newborn. There is no meterstick of human values. Aristotle was among the first to notice that human values are incommensurable. You can’t, he argued, measure the true (as opposed to market) value of beds and shoes on the same scale of value. AI supposes otherwise.

AI also assumes that in a decision, there are only two possibilities: one option is better than the other, in which case you should choose it, *or* they’re equally good, in which case you can just flip a coin. Hard choices suggest otherwise. When you are agonising between two careers, neither is better than the other, but they aren’t equally good either – they are simply different. The values that govern hard choices allow for more possibilities: options might be on a par*.*Many of our choices between jobs, people to marry, and even government policies are on a par. AI architecture currently makes no room for such hard choices.

Finally, AI presumes that the values in a choice are “out there” to be found. But sometimes we create values through the very process of making a decision. In choosing between careers, how much does financial security matter as opposed to work satisfaction? You may be willing to forgo fancy meals in order to make full use of your artistic talents. But I want a big house with a big garden, and I’m willing to spend my days in drudgery to get it.

Our value commitments are up to us, and we create them through the process of choice. Since our commitments are internal manifestations of our will, observing our behaviour won’t uncover their specificity. AI, as it is currently built, supposes values can be programmed as part of a reward function that the AI is meant to maximise. Human values are more complex than this.

\*\*\*

So where does that leave us? There are three possible paths forward.

Ideally, we would try to develop AI architecture that respects the incommensurable, parity-tolerant and self-created features of human values. This would require serious collaboration between computer scientists and philosophers. If we succeed, we could safely outsource many of our decisions to machines, knowing that AI will mimic human decision-making at its best. We could prevent AI from taking over the world while still allowing it to transform human life for the better.

If we can’t get AI to respect human values, the next best thing is to accept that AI should be of limited use to us. It can still help us crunch numbers and discern patterns in data, operating as an enhanced calculator or smart phone, but it shouldn’t be allowed to make any of our decisions. This is because when an AI makes a decision, say, to swerve your car to avoid hitting a fox, at some risk to your life, it’s not a decision made on the basis of human values but of alien, AI values. We might reasonably decide that we don’t want to live in a world where decisions are made on the basis of values that are not our own. AI would not take over the world, but nor would it fundamentally transform human life as we know it.

The most perilous path – and the one towards which we are heading – is to hope in a vague way that we can strike the right balance between the risks and benefits of AI. If the mismatch between AI architecture and human values is beyond repair, we might ask ourselves: how much risk of annihilation are we willing to tolerate in exchange for the benefits of allowing AI to make decisions for us, while, at the same time, recognising those decisions will necessarily be made based on values that are not our own?

*That*decision, at least, would be one made by us on the basis of our human values. The overwhelming likelihood, however, is that we get the trade-off wrong. We are, after all, only human. If we take this path, AI *could* take over the world. And it would be cold comfort that it was our human values that allowed it to do so.

*Ruth Chang is the Chair and Professor of Jurisprudence at the University of Oxford and a Professorial Fellow at University College, Oxford. She is the author of*[Hard Choices](https://www.cambridge.org/core/journals/journal-of-the-american-philosophical-association/article/abs/hard-choices/A4BEB486621AAEFEC50145087501D584)*and a*[*TED talk*](https://www.ted.com/talks/ruth_chang_how_to_make_hard_choices?language=en)*on decision-making.*

*This article is part of the*[*Agora*](https://www.newstatesman.com/2019/02/agora-market-place-ideas)*series, a collaboration between the*New Statesman*and Aaron James Wendland, Senior Research Fellow in Philosophy at Massey College, Toronto. He tweets*[*@aj\_wendland*](https://twitter.com/aj_wendland)*.*