Stop the Robot Apocalypse

Amia Srinivasan

Doing Good Better:
Effective Altruism and a Radical New Way to Make a Difference
by William MacAskill.
Guardian Faber, 325 pp., £14.99, August, 978 1 78335 049 0

Philosophy, Wittgenstein said, 'leaves everything as it is'. It sounds like a complaint, but actually it was a recommendation. Philosophy at its best, Wittgenstein thought, resists the scientific impulse to treat the world as a theoretical construct. It is not a view shared in the main by contemporary philosophers. What is philosophy supposed to do if not theorise? At the same time most philosophers are happy to leave everything as it is in a more prosaic sense: that is, by not really changing anything. Philosophers may talk about justice or rights, but they don't often try to reshape the world according to their ideals. Maybe that's for the best. Philosophers have a tendency to slip from sense into seeming absurdity: a defence of abortion ends up defending infanticide; an argument for vegetarianism turns into a call for the extermination of wild carnivores.

A new generation of moral philosophers is determined to break with this tradition of ineffectuality. The goal of the 'effective altruists' is not only to theorise the world, but to use their theories to leave the world a better place than they found it. Their leader is William MacAskill, a 28-year-old lecturer at Oxford. As graduate students MacAskill and his friend Toby Ord committed themselves to donate most of their future earnings to charity (in MacAskill's case anything above £20,000, in Ord's £18,000), and set themselves the task of figuring out how to make best use of the money they had pledged. The result was Giving What We Can, a charity that encourages people to hand over at least 10 per cent of their future incomes for philanthropic purposes, and advises them on how to get the most out of their money. Since the charity was founded in 2009 it has received more than $400 million in pledges, much of it from young philosophers. In 2011, MacAskill set up 80,000 Hours (the name refers to the number of hours the average person works over a lifetime), a charity that helps people make career choices with the aim of maximising social benefit; it raised eyebrows early on by advising graduates to become philanthropic bankers rather than NGO workers. The two organisations are incorporated as the Centre for Effective Altruism, based in Oxford, and are in the van of a global movement, encompassing groups such as GiveWell (founded by two hedge-fund managers at around the same time as MacAskill and Ord started their work), The Life You Can Save (founded by the philosopher Peter Singer), Good Ventures (founded by the Facebook cofounder Dustin Moskovitz and his wife, Cari Tuna, who have pledged to give away most of their money), Animal Charity Evaluators (an 80,000 Hours spin-off) and the Open Philanthropy Project (a collaboration between GiveWell and Good Ventures).

In *Doing Good Better*, MacAskill sets out the thinking behind effective altruism. His main claim, familiar from the utilitarian tradition out of which the movement emerges, is that we should seek not only to do good, but to do the most good we can. To do that we need empirical research – research his organisations provide – into the amount of good created by various different charities, types of consumption, careers and so on. MacAskill proposes that 'good', here, can be understood roughly in terms of quality-adjusted life-years (Qalys), a unit that allows welfare economists to compare benefits of very different sorts. One Qaly is a single year of life lived at 100 per cent health. According to a standardised scale, a year as an Aids patient not on antiretrovirals is worth 0.5 Qalys; a year with Aids lived on antiretrovirals is worth 0.9 Qalys. A year of life for a blind person is worth 0.4 Qalys; a year of life as a non-blind, otherwise healthy person is worth 1 Qaly. (These numbers are based on self-reporting by Aids patients and blind people, which raises some obvious worries. For example, dialysis patients rate their lives at 0.56 Qalys – significantly higher than the 0.39 Qalys predicted by people who don't need dialysis. Maybe this is because dialysis isn't as bad as we think. Or maybe it's because dialysis is so awful that you forget just how much better your life was without it.) To calculate whether, given the choice, it would be better to cure the blind person or improve the life of the Aids patient, you must take into account the increase in both life quality and life expectancy that would be caused by the intervention. Giving a 40-year-old Aids patient antiretrovirals would give her a 40 per cent jump in life quality (from 50 per cent to 90 per cent) for five years, and also would give her an additional five years of life at 90 per cent, for a total of 6.5 Qalys ((0.4 x 5) + (0.9 x 5) = 6.5). Curing a blind 20-year-old, assuming he lives to be 70, would increase his quality of life from 40 per cent to 100 per cent for 50 years, for a total of 30 Qalys. So curing the blind person is more valuable, in terms of Qalys, than giving drugs to the Aids patient. Thinking in terms of Qalys makes it possible to compare that which seemingly cannot be compared: blindness with Aids; increases in life expectancy with increases in life quality. Qalys free us from the specificity of people's lives, giving us a universal currency for misery.

However, when deciding what to do – what job to get, which charity to donate to, whether to buy Fairtrade or not – it isn't enough, according to MacAskill, to think in terms of Qalys. We must also think both marginally and counterfactually. The idea that value should be measured on the margin is familiar from economics; it's what explains the fact that, say, heating repairmen make more money than childcare workers. Presumably childcare workers produce more total value than heating repairmen, but because the supply of childcare workers is greater than the supply of good repairmen, we will pay more for an additional repairman than an additional childcare worker. The average value of a childcare worker might be higher than the average value of a repairman, but the repairman has the greater *marginal* value. (Another way of putting this: coffee might be really important to you, but if you've already had three cups you're probably not going to care as much about a fourth.) Similarly, MacAskill reasons, when giving money to charity or deciding, say, whether to become a doctor, we should focus on marginal rather than average value. The average doctor in the developed world helps save a lot of lives, but the marginal doctor – because the supply of doctors is large, and most of the life-saving work is already covered – doesn't. The marginal doctor in the developing world has greater value, since the supply of doctors is lower there. MacAskill estimates that a doctor practising in a very poor country adds about a hundred times as much marginal value (measured in Qalys) as a doctor practising in the UK. (In general, MacAskill says, a pound spent in a poor country can do one hundred times more good than it can in a rich one, a heuristic he calls the '100x Multiplier'.)

But before signing up for a medical career in sub-Saharan Africa you should be careful, MacAskill warns, to evaluate the counterfactual. That is, you have to ask yourself what would happen if you didn't become a doctor at all. Let's say that as a doctor in the developing world you'd save the equivalent of 300 lives (or 10,950 Qalys, at an average 36.5 Qalys per life) over a 40-year career. Yet if you didn't take that job, someone else probably would; they may not save quite as many lives as you, but they would save most of them. Meanwhile you could quit medicine, take a high-paying finance job and donate most of your salary each year to the most effective charities. MacAskill estimates that you can expect to save a life with $3400 by donating to the Against Malaria Foundation, which provides insecticide-coated bed nets to poor families. A financier who worked for 40 years, donating $50,000 a year to the Against Malaria Foundation, could expect to save around 580 lives – lives, significantly, that would not have been saved otherwise. Qaly thinking frees us from considering the specificity of whom we are helping; marginal and counterfactual thinking frees us from the specificity of ourselves. What matters isn't who does the good, only that good is done.

But don't many lucrative careers have bad social effects? Up until recently MacAskill argued that such effects were morally irrelevant, again by counterfactual reasoning: if you didn't take the banking job someone else would, so the harm would be done anyway. (In an academic paper published last year, he compares a philanthropic banker to Oskar Schindler, who provided munitions to the Nazis as a means of saving the lives of 1200 Jews; if Schindler hadn't manufactured the arms, some other Nazi would have, without saving any Jewish lives.) More recently MacAskill and his team at

80,000 Hours have backed away from this 'replaceability thesis', conceding that it's harder than they initially thought to evaluate the counterfactuals. For example, there's good economic reason to think that going into banking really does increase the total number of bankers, and doesn't simply change who does the banking. MacAskill says he no longer recommends that people go into banking, or at least not the parts of it that he thinks cause direct harm: creating risks that will be borne by unsuspecting taxpayers, or selling products that no properly informed person would buy. Instead 80,000 Hours now encourages people to take what it sees as morally neutral or positive jobs: quantitative hedge-fund trading, management consulting, technology start-ups. (You can take a careers quiz on the 80,000 Hours website; I was told to become a consultant, because of its earning-to-give potential and the general business education it provides. When I changed my answers to say that I was bad at maths I was told to go into politics.)

The results of all this number-crunching are sometimes satisfyingly counterintuitive. Deworming has better educational outcomes among Kenyan schoolchildren than increasing the numbers of textbooks or teachers. If you want to improve animal welfare, it's better to stop eating eggs than beef, since caged layer hens live worse lives than farmed cows, and because eating eggs consumes more animals than eating beef: the average American consumes 0.8 layer hens but only 0.1 beef cows per year. Buying Fairtrade goods can be worse than buying regular goods, since the extra cost goes mostly to middlemen rather than farmers, and when it doesn't, it benefits farmers in relatively rich countries: because Fairtrade standards are hard to meet, most Fairtrade coffee production comes from Mexico and Costa Rica rather than, say, Ethiopia, where the marginal pound would go much further. The green value of buying locally grown food is overblown, too, since transport accounts for only 10 per cent of the carbon footprint of food, while 80 per cent of it is generated in production; tomatoes grown in the UK can have five times the carbon footprint of tomatoes shipped from Spain because of the energy required to hothouse them. If you're really committed to minimising your carbon footprint, MacAskill recommends donating to the carbon offsetting charity Cool Earth; he estimates that the average American could offset all his carbon emissions by donating $105 a year. There isn't much point in unplugging your electricals, either: leaving your mobile phone charger plugged in for a whole year contributes less to your carbon footprint than one hot bath.

Doing Good Better is a feel-good guide to getting good done. It doesn't dwell much on the horrors of global inequality, and sidesteps any diagnosis of its causes. The word 'oppression' appears just once. This is surely by design, at least in part. According to MacAskill's moral worldview, it is the consequences of one's actions that really matter, and that's as true of writing a book as it is of donating to charity. His patter is calculated for maximal effect: if the book weren't so cheery, MacAskill couldn't expect to inspire as much do-gooding, and by his own lights that would be a moral failure. (I'm not saying it doesn't work. Halfway through reading the book I set up a regular donation to GiveDirectly, one of the charities MacAskill endorses for its proven efficacy. It gives unconditional direct cash transfers to poor households in Uganda and Kenya.)

But the book's snappy style isn't just a strategic choice. MacAskill is evidently comfortable with ways of talking that are familiar from the exponents of global capitalism: the will to quantify, the essential comparability of all goods and all evils, the obsession with productivity and efficiency, the conviction that there is a happy convergence between self-interest and morality, the seeming confidence that there is no crisis whose solution is beyond the ingenuity of man. He repeatedly talks about philanthropy as a deal too good to pass up: 'It's like a 99 per cent off sale, or buy one, get 99 free. It might be the most amazing deal you'll see in your life.' There is a seemingly unanswerable logic, at once natural and magical, simple and totalising, to both global capitalism and effective altruism. That he speaks in the proprietary language of the illness – global inequality – whose symptoms he proposes to mop up is an irony on which he doesn't comment. Perhaps he senses that his potential followers – privileged, ambitious millennials – don't want to hear about the iniquities of the system that has shaped their worldview. Or perhaps he thinks there's no irony here at all: capitalism, as always, produces the means of its own correction, and effective altruism is just the latest instance.

Yet there is no principled reason why effective altruists should endorse the worldview of the benevolent capitalist. Since effective altruism is committed to whatever would maximise the social good, it might for example turn out to support anti-capitalist revolution. And although MacAskill focuses on health as a proxy for goodness, there is no principled reason, as he points out, why effective altruism couldn't also plug values like justice, dignity or self-determination into its algorithms. (There's also no reason why one couldn't 'earn to give' to help radical causes; Engels worked at a mill in Manchester to support Marx's writing of *Capital*.) Effective altruism has so far been a rather homogenous movement of middle-class white men fighting poverty through largely conventional means, but it is at least in theory a broad church. Indeed one element of the movement is turning its attention towards what members like to call 'systemic change', taking up political advocacy on issues ranging from factory farming to immigration reform. Even in these cases, the numbers are what matter. MacAskill describes how he helped an Oxford PPE student work out whether or not she should get into electoral politics. He calculates that historically, the odds of a politically ambitious Oxford PPE student becoming an MP have been one in thirty (he notes that this reflects 'some disappointing facts about political mobility and equal representation in the UK'). Applying some conservative estimates of the resources an average MP gets to control, he prices the marginal expected value of the student's running for Parliament at £8 million, which turns out to be high enough, compared with the expected value of other careers she might pursue, to justify the move into politics. It's not clear that anyone with less conventional political ambitions would get the same pass. What's the expected marginal value of becoming an anti-capitalist revolutionary? To answer that you'd need to put a value and probability measure on achieving an unrecognisably different world – even, perhaps, on our becoming unrecognisably different sorts of people. It's hard enough to quantify the value of a philanthropic intervention: how would we go about quantifying the consequences of radically reorganising society?

MacAskill seems to think there is no moral calculation that can't be made to fit on the back of his envelope; any uncertainty we might have about precise values or probabilities can be priced into the model. (His doctoral dissertation was on the modelling of moral uncertainty.) But the more uncertain the figures, the less useful the calculation, and the more we end up relying on a commonsense understanding of what's worth doing. Do we really need a sophisticated model to tell us that we shouldn't deal in subprime mortgages, or that the American prison system needs fixing, or that it might be worthwhile going into electoral politics if you can be confident you aren't doing it solely out of self-interest? The more complex the problem effective altruism tries to address – that is, the more deeply it engages with the world as a political entity – the less distinctive its contribution becomes. Effective altruists, like everyone else, come up against the fact that the world is messy, and like everyone else who wants to make it better they must do what strikes them as best, without any final sense of what that might be or any guarantee that they're getting it right.

More worrying than the model's inability to tell us anything very useful once we move outside the circumscribed realm of controlled intervention is its susceptibility to being used to tell us exactly what we want to hear. A three-day conference, 'Effective Altruism Global', was held this summer at Google's headquarters in Mountain View, California. While some of the sessions focused on the issues closest to MacAskill's heart – cost-effective philanthropy, global poverty, career choice – much of it was dominated, according to Dylan Matthews, who was there and wrote about it for *Vox*, by talk of existential risks (or x-risks, as the community calls them). An x-risk, as defined by the Oxford philosopher Nick Bostrom, who popularised the concept, is an event that would 'permanently and drastically curtail humanity's potential' – total annihilation is the obvious case. Given the number of people who might live in the future if not for such an event – Bostrom estimates the figure at 10^{52}, assuming that we master interstellar travel and the uploading of human minds to computers – the expected value of preventing an x-risk dwarfs the value of, say, curing cancer or preventing genocide. This is so even if the probability of being able to do anything about an x-risk is vanishingly small. Even if Bostrom's 10^{52} estimate has only a 1 per cent chance of being correct, the expected value of reducing an x-risk by one billionth of one billionth of a percentage point (that's 0.0000000000000000001 per cent) is still a hundred billion times greater than the value of saving the lives of a billion people living now. So it turns out to be better to try to prevent some hypothetical x-risk, even with an extremely remote chance of being able to do so, than to help actual living people.

X-risks could take many forms – a meteor crash, catastrophic global warming, plague – but the one that effective altruists like to worry about most is the 'intelligence explosion': artificial intelligence taking over the world and destroying humanity. Their favoured solution is to invest more money in AI research. Thus the humanitarian logic of effective altruism leads to the conclusion that more money needs to be spent on computers: why invest in anti-malarial nets when there's a robot apocalypse to halt? It's no surprise that effective altruism is popular in Silicon Valley: PayPal founder Peter Thiel, Skype developer Jaan Tallinn and Tesla CEO Elon Musk are all major financial supporters of x-risk research.* Who doesn't want to believe that their work is of overwhelming humanitarian significance?

The subtitle of *Doing Good Better* promises 'a radical new way to make a difference'; one of the organisers of the Googleplex conference declared that 'effective altruism could be the last social movement we ever need.' But effective altruism, so far at least, has been a conservative movement, calling us back to where we already are: the world as it is, our institutions as they are. MacAskill does not address the deep sources of global misery – international trade and finance, debt, nationalism, imperialism, racial and gender-based subordination, war, environmental degradation, corruption, exploitation of labour – or the forces that ensure its reproduction. Effective altruism doesn't try to understand how power works, except to better align itself with it. In this sense it leaves everything just as it is. This is no doubt comforting to those who enjoy the status quo – and may in part account for the movement's success.

Yet behind MacAskill's cheery exhortation to invest in anti-malarial nets lies a moral philosophy that really is radical. In 1972 Peter Singer published his paper 'Famine, Affluence and Morality', a classic of contemporary utilitarianism, in which he compares a Westerner who spends money on luxuries rather than donating it to the developing world to someone who walks by a drowning child rather than get his clothes muddy. We can all agree that the second case is morally abhorrent, but not everyone has the same qualms about the first. What's the difference? Does it really matter, Singer asks, that a child in the developing world is thousands of miles away rather than in front of us? If not, then all of us who don't merely subsist, who spend money on ourselves when it would be worth significantly more to someone else, are morally implicated in murder, just like the man who allows the child to drown. The vast scale of global inequality – if your income is more than £34,000 per year, adjusted for purchasing power, you're in the global 1 per cent – means that even the smallest luxuries (going to the cinema, a second pair of shoes, a drink at the pub) may be morally unacceptable.

Effective altruism takes up the spirit of Singer's argument but shields us from the full blast of its conclusion; moral indictment is transformed into an empowering investment opportunity. Instead of downgrading our lives to subsistence levels, we are encouraged to start with the traditional tithe of 10 per cent, then do a bit more each year. Thus effective altruism dodges one of the standard objections to utilitarianism: that it asks too much of us. But it isn't clear how the dodge is supposed to work. MacAskill tells us that effective altruists – like utilitarians – are committed to doing the most good possible, but he also tells us that it's OK to enjoy a 'cushy lifestyle', so long as you're donating a lot to charity. Either effective altruism, like utilitarianism, demands that we do the most good possible, or it asks merely that we try to make things better. The first thought is genuinely radical, requiring us to overhaul our daily lives in ways unimaginable to most. (Singer repeats his call for precisely such an overhaul in his recent book *The Most Good You Can Do*, and Larissa MacFarquhar's *Strangers Drowning* is a set of portraits of 'extreme altruists' who have answered the call.†) The second thought – that we try to make things better – is shared by every plausible moral system and every decent person. If effective altruism is simply in the business of getting us to be more effective when we try to help others, then it's hard to object to it. But in that case it's also hard to see what it's offering in the way of fresh moral insight, still less how it could be the last social movement we'll ever need.

A more pressing objection to utilitarianism is not that it demands too much, but that it demands the wrong things, the things that constitute us as humans: our personal attachments, loyalties and identifications. On the utilitarian view, a pound spent without maximal effect is a pound spent immorally. Luxuries are naturally ruled out, but so is spending on worthwhile causes to which you might feel some personal affinity. Here MacAskill agrees: to choose to donate to a relatively cost-ineffective charity just because it's close to your heart – the local soup kitchen, or a seeing-eye dog charity in honour of a blind relative (it costs £32,400 to train one seeing-eye dog and its owner) – is wrong. How far should the effective altruist go with this logic? If you're faced with the choice between spending a few hours consoling a bereaved friend, or earning some money to donate to an effective charity, the utilitarian calculus will tell you to do the latter. If effective altruists really are committed to doing the most good, they should say the same. If however they are merely committed to doing a lot of good, then they will say that you can stay with your friend so long as you're doing sufficient good elsewhere. But even this more moderate view misconceives the situation. You should stay and console your friend not because you've already met your do-gooding quota, but because it's *your* friend that is in distress. This is also the reason you shouldn't deal in subprime mortgages or make money from the exploitation of labour, even if the good effects would outweigh the bad: it's *your* life, and it matters, morally speaking, what you do with it, and not just – as MacAskill suggests – what is done because of it.

That emphasis on 'your' is something that utilitarians often find conceptually mystifying, or at least a moral distraction. Here, for example, is MacAskill talking about his visit to the Hamlin Fistula Hospital in Addis Ababa, and his later decision not to donate to its main charitable benefactor:

> I'd hugged the women who suffered from this condition, and they'd thanked me for visiting them. It had been an important experience for me: a vivid first-hand demonstration of the severity of the problems in the world. This was a cause I had a personal connection with. Should I have donated to the Fistula Foundation, even knowing I could do more to help people if I donated elsewhere? I do not think so. If I were to give to the Fistula Foundation rather than to charities I thought were more effective, I would be privileging the needs of some people over others for emotional rather than moral reasons. That would be unfair to those I could have helped more. If I'd visited some other shelter in Ethiopia, or in any other country, I would have had a different set of personal connections. It was arbitrary that I'd seen this particular problem at close quarters.

That word 'arbitrary' is striking. It is indeed arbitrary that MacAskill went to this hospital and not another, in Ethiopia and not some other country, just as it is arbitrary that we have the family, friends, lovers and neighbours we do. But doesn't such arbitrariness come to mean something else, ethically speaking, when it is constitutive of our personal experience: when it becomes embedded in the complex structure of commitments, affinities and understandings that comprise social life? We might even think that the arbitrariness of time and place is transformed into something else, ethically speaking, through the exchange of a fleeting hug or thanks. What's more, MacAskill's talk of fairness is too easy. It is no doubt unfair that some of the world's worst off are helped while others aren't. But isn't it just as unfair that the Ethiopian women MacAskill met are victims of a debilitating condition that is too costly to be 'worth' funding? And what of the victims of austerity or rising inequality in the first world? MacAskill's reminder that these people are still among the world's richest is cold comfort (it also obscures what all those trampled by the ruling class everywhere may have in common).

When MacAskill says that helping the Ethiopian women he met would be 'arbitrary' and 'unfair', he means to speak from what the 19th-century utilitarian Henry Sidgwick called 'the point of view of the universe'. But in so doing MacAskill is trying to step outside what is unavoidably the scene of ethical action: one's own point of view. MacAskill thinks this self-transcendence – or as close as we non-saints can get to it – is essential if we are going to meet the ethical demands of our day. Wittingly or not, he believes, we are all like A&E doctors, forced to perform triage lest more people suffer and die than have to. What is required is impersonal, ruthless decision-making, heart firmly reined in by the head. This is not our everyday sense of the ethical life; such notions as responsibility, kindness, dignity and moral sensitivity will have to be radically reimagined if they are to survive the scrutiny of the universal gaze. But why think this is the right way round? Perhaps it is the universal gaze that cannot withstand our ethical scrutiny.

There is a small paradox in the growth of effective altruism as a movement when it is so profoundly individualistic. Its utilitarian calculations presuppose that everyone else will continue to conduct business as usual; the world is a given, in which one can make careful, piecemeal interventions. The tacit assumption is that the individual, not the community, class or state, is the proper object of moral theorising. There are benefits to thinking this way. If everything comes down to the marginal individual, then our ethical ambitions can be safely circumscribed; the philosopher is freed from the burden of trying to understand the mess we're in, or of proposing an alternative vision of how things could be. The philosopher is left to theorise only the autonomous man, the world a mere background for his righteous choices. You wouldn't be blamed for hoping that philosophy has more to give. □