Journal of Practical Ethics

Volume 7, Number 1. June 2019

CONTENTS

The Journal of Practical Ethics is available online, free of charge, at:
http://jpe.ox.ac.uk

Editorial Policy
The *Journal of Practical Ethics* is an invitation only, blind-peer-reviewed journal. It is entirely open access online, and print copies may be ordered at cost price via a print-on-demand service. Authors and reviewers are offered an honorarium for accepted articles. The journal aims to bring the best in academic moral and political philosophy, applied to practical matters, to a broader student or interested public audience. It seeks to promote informed, rational debate, and is not tied to any one particular viewpoint. The journal will present a range of views and conclusions within the analytic philosophy tradition. It is funded through the generous support of the *Uehiro Foundation in Ethics and Education*.

Copyright

ISSN: 2051-655X

Being Good in a World of Need: Some Empirical Worries and an Uncomfortable Philosophical Possibility

LARRY S. TEMKIN

Rutgers University

ABSTRACT

In this article, I present some worries about the possible impact of global efforts to aid the needy in some of the world's most desperate regions. Among the worries I address are possible unintended negative consequences that may occur elsewhere in a society when aid agencies hire highly qualified local people to promote their agendas; the possibility that foreign interests and priorities may have undue influence on a country's direction and priorities, negatively impacting local authority and autonomy; and the related problem of outside interventions undermining the responsiveness of local and national governments to their citizens.

Another issue I discuss is the possibility that efforts to aid the needy may involve an *Each-We Dilemma*, in which case conflicts may arise between what is *individually* rational or moral, and what is *collectively* rational or moral. Unfortunately, it is possible that if *each* of us does what we have *most* reason to do, morally, in aiding the needy, we *together* will bring about an outcome which is *worse*, morally, in terms of its overall impact on the global needy.

The article ends by briefly noting a number of claims and arguments that I made in my 2017 Uehiro Lectures regarding how good people should respond in a world of need. As I have long argued, I have no doubt that those who are well off are open to serious moral criticism if they ignore the plight of the needy. Unfortunately, however, for a host of both empirical and philosophical reasons, what one should do in light of that truth is much more complex, and murky, than most people have realized.

PART I. INTRODUCTION.

For most of my life, I have been deeply concerned about the problems of the global needy, and for many years, I have published and lectured on the topic (Temkin 1999, 2004a, 2004b).

Along with Peter Singer, I helped launch the University of Manchester chapter of the Effective Altruist organization, *Giving What We Can* and, along with Jeffrey Sachs, I helped launch the Princeton University chapter of that organization. In my lectures and writings, I have long contended that most of those in the developed world are open to serious moral criticism, when they basically ignore, as most of us do, the plight of the world's needy. I continue to think that. Nevertheless, philosophers are required to subject even their deepest and most longstanding commitments to critical scrutiny, and to follow the arguments wherever they lead. And in recent years, I have become increasingly worried about possible negative impacts of global efforts to aid the needy in some of the world's most desperate regions. In this article, I raise some of those worries.

The article contains four main parts. In Part II, I address some worries about certain marketplace distortions that can arise as a result of aid efforts on behalf of the needy in some of the world's poorest countries. In particular, I note some possible unintended negative consequences that may occur elsewhere in a society when aid agencies hire highly qualified local people to promote their agendas. In Part III, I discuss the possibility that foreign interests and priorities may have undue influence on a country's direction and priorities, negatively impacting local authority and autonomy; and the related problem of outside interventions undermining the responsiveness of local and national governments to their citizens. In Part IV, I discuss the possibility that efforts to aid the needy may involve an *Each-We Dilemma*, in which case conflicts may arise between what is *individually* rational or moral, and what is *collectively* rational or moral. Drawing on results from my book, *Rethinking the Good* (2012), I argue that it is possible that if *each* of us does what we have *most* reason to do, morally, in aiding the needy, we *together* will bring about an outcome which is *worse*, morally, in terms of its overall impact on the global needy. In Part V, I respond to the view that we shouldn't provide direct aid to people in the world's poorest countries,

because doing so contributes to poor governance within such countries. In Part VI, I end by offering a few claims that I made in my 2017 Uehiro Lectures regarding how good people should respond in a world of need. Unfortunately, in this article I must be content to merely offer those claims, without further argument.

I am acutely aware that this article, which is based on the third of my three 2017 Uehiro Lectures, is only a preliminary treatment of the issues explored. Equally important, there are a host of other crucial issues related to the topic of how good people should respond in a world of need that this article doesn't even broach. But no one article can address every important issue, and I believe the issues I am addressing here are deserving of much more attention than they have typically been given, at least in the philosophical literature. I hope to give a much fuller treatment of this important topic in a book tentatively titled *Being Good in a World of Need* to be published as part of the Uehiro Lectures Book Series.

Because this article mainly raises worries about the possible negative effects of efforts to aid the needy in some of the world's poorest countries, let me emphasize, at the outset, that I remain as committed, as ever, to the view that those of us in a position to do so—which includes almost everyone in the so-called developed world—have a strong moral imperative to find ways of effectively helping our world's worst-off members. Unfortunately, however, for a host of both empirical and philosophical reasons, it is much less clear to me now, than it once was, what we should actually do in light of that truth.

PART II. MARKET PLACE DISTORTIONS.

One common worry regarding global aid, concerns the possibility of corruption, and ways in which aid efforts may benefit evil agents, and give rise to perverse incentives and indirect, negative effects (Easterly 2006, Moyo 2010, Wenar 2011, Deaton 2013, and Temkin 2017b). Unfortunately, aid efforts can also give rise to indirect, negative effects when *no* corruption or evil actions are involved. Moreover, these negative effects are easily overlooked and difficult to quantify.

I start with a point familiar to global health experts. International aid groups promote many worthy projects. They might improve the water supply, build new schools, bring electricity to villages, construct medical clinics, and so on. As a result, they may hire many local workers: managers, engineers, principals, teachers, doctors, nurses, administrative staff, drivers, road pavers, well diggers, wire stringers, and

so on. Naturally, the most *effective* aid groups seek to hire the *best* people they can for these positions. Ideally, they will hire personable people with good leadership, managerial, and communication skills, who work well with others, and are dedicated, trustworthy, hardworking, reliable, and so on.

Aid groups will be in competition with each other for such people and, thanks to their donors, will be able to offer higher pay, fewer hours, and better working conditions than the local standard. Accordingly, highly qualified people from across the region will seek these jobs.

As described, so far, this sounds like a win/win/win situation. Hiring such talented people will be great for the needy, the workers themselves, and the aid groups, enabling them to truthfully show their donors how they have effectively achieved their goals. Unfortunately, left out of this rosy account is what happens elsewhere in the system as a result of the successful, well-intentioned, aid efforts.

In particular, one needs to worry about the indirect effects of hiring such people away from whatever jobs they might otherwise occupy (Leif Wenar also recognizes this worry, see 2011). Governments in poor countries desperately need talented engineers, accountants, lawyers, teachers, doctors, nurses, managers, and civil servants working on behalf of the general public. Unfortunately, however, most poor governments cannot match the pay scale or working conditions that many aid groups offer. This may result in an internal "brain and character drain" away from public sector jobs, which may have a significantly deleterious effect on the efficiency and success of the government, the economy, and public projects.

Moreover, depending on the disparities between salaries and working conditions, one might see highly trained professionals leaving jobs that require all of their talents, for jobs for which they are overqualified. Thus, due to marketplace distortions that well-funded aid groups may inadvertently create, some outstanding teachers, engineers, accountants, lawyers, doctors, nurses, and civil servants may happily give up their posts to become administrators, clerks, drivers, or manual laborers. If that happens, the overall costs their society bears, when such people are no longer performing jobs befitting their talents, may substantially outweigh the relative gains their society gets due to their successfully fulfilling their new positions.

Here is a related problem. Highly talented, hard-working people of great character will always be in demand. Such people may well get used to an aid group's pay scale and high quality working conditions. Moreover, such people may receive special training or make connections with well-placed aid officials which enable them

to compete for comparable positions outside of their countries. So, what will happen when an aid group shuts down its local operation? Its highly talented workforce *might* return to the low pay and poor working conditions of their previous places of employment, where their skills might be desperately needed. Or, they *might* seek better prospects in the developed world where the need for their talents is much less great, but the personal rewards are far greater.

There is an old question: "how do you keep people down on the farm once they have seen the lights of the big city?" There is a kernel of truth embedded in that question which underlies my worry here. No one can blame aid groups for hiring highly talented people to efficiently promote their important goals. Nor can one blame such people for bettering themselves and their families. Yet, together, these perfectly understandable and laudable goals may contribute to both internal and external "brain and character drains" that can be deeply problematic for the world's poorest countries (though, importantly, so-called brain "drains" can also have positive effects on poor countries when accompanied by remittances that overseas workers send back to their home countries).

Thus, an aid group's gains, which are often readily identified and quantified, may be offset by indirect losses elsewhere in the system in ways that are easily overlooked and difficult to quantify. This can result in a distorted picture of the overall good that an aid group is doing. I am not claiming that the net effect of such tradeoffs will necessarily be negative, though in some cases it may be. But merely that the desirability of supporting an aid group must take *full* account of the opportunity costs of doing so, including not only where else I might spend my money, but also what else an aid group's local workers would be doing, if they weren't working for the aid group. Unfortunately, given the countless aid groups that operate in some of the world's poorest regions, the *cumulative* impact of the negative effects that I have been describing may be substantial.

PART III. RESPONSIVENESS AND THE IMPORTANCE OF GOOD GOVERNANCE.

Let me turn next to a worry raised by Angus Deaton, a leading international development expert and the 2015 Nobel Prize Winner in Economics. Deaton is deeply concerned about the world's needy, but after analyzing data about economic development in the world's poorest regions, and searching for correlations between how

much international aid a poor country receives and its level of social and economic development, Deaton has arrived at a striking conclusion: people like Peter Singer are doing more harm than good! Specifically, Deaton believes that if we genuinely want to aid the world's needy, we must find some route to do so other than by contributing to aid groups that work directly in the world's poorest regions to ameliorate their desperate conditions (see Deaton 2013, Chapter Seven).

Deaton knows that his conclusion is at odds with what most people think. It is, after all, deeply counterintuitive to believe that if external funding pours into a region of great need, explicitly earmarked to address those needs, that, overall, the result should prove fruitless, at best, or harmful, at worst. Deaton also recognizes that fully explaining his findings is not easy. Still, Deaton suggests several factors that might help account for his findings and support his counterintuitive conclusion.

Consider first,

> *The Paradox of Aid*: in countries where the need is greatest, aid won't help; while in countries where aid would help most, it isn't needed.

If there is a kernel of truth to this Paradox, as many development economists believe, it reflects the crucial role that governments play in their countries' social and economic progress. The basic thought is that good governments find a way to take care of their people's basic needs; while poor governments are either unable, or unwilling, to do so. Even worse, poor governments tend to obstruct aid efforts, so that any gains will be short term, at best. On this view, substantial and long-lasting social and economic gains require a well-functioning government which can formulate and effectively implement plans to develop infrastructure, energy, food production, schools, the health system, etc. Aid groups, no matter how well-intentioned or well-funded, cannot accomplish this on their own.

What makes a government well-functioning? Deaton believes the key component of a well-functioning government is that it be responsive to its citizens' needs, interests, and will. With that in mind, Deaton suggests that what primarily accounts for the counterproductiveness of international aid efforts in the world's poorest regions is that they tend to undermine the governments' responsiveness to their citizens. But, to repeat, on Deaton's view, it is precisely such responsiveness that is necessary for any poor country's long-term social and economic development.

The mechanisms by which international aid may undermine governments' re-

sponsiveness to their people include the following. First, corrupt governments may find ways of capturing aid resources for their own purposes. They may impose licensing fees that fill their coffers; tax or demand kickbacks from aid beneficiaries; extort bribes in return for government cooperation; insist that aid groups employ their supporters; require aid groups to supply them with food, medicine, or other supplies; and so on. In sum, there are many ways in which corrupt governments can divert aid resources to strengthen their positions and advance their agendas. This can enable such governments to be indifferent and unresponsive to their citizens' needs, interests, and will, and to put their own interests, and those of their supporters, ahead of their general populations'.

Second, in many of the world's poorest regions many outside aid groups operate. Some address hunger; others poverty; others rape or victims of sectarian violence; others victims of particular illnesses, such as malaria, tuberculosis, diarrhea, or AIDs; others pre-natal, post-natal, and maternal health care; others health care more generally; others education; others female empowerment; others infrastructure; and so on. Of course, some agencies will address multiple concerns. This all sounds desirable. But if *so* many aid groups are helping in so many ways, how come the problems of the needy continue to persist in the world's poorest regions? Is it merely because not *enough* aid groups have been involved? Or not *enough* resources efficiently spent to eradicate the problems?

Deaton has another hypothesis. He believes that with so many aid groups working to help the needy, local governments can abdicate their responsibilities to provide for their citizens' basic needs, and leave that task to the aid groups. The local governments can then shift the blame for any unmet needs to the aid groups, who have failed to fully deliver on their promises to help the needy! In other words, the well-intentioned interventions by aid groups can undermine the local governments' responsiveness to their citizens' needs, interests, and will. But, of course, if Deaton is correct, such responsiveness is a *key* characteristic of good governance, without which there can be no *hope* of a *lasting* solution to the social and economic woes of the world's poorest nations.

The preceding points are intimately related to a third point. Generally, effective governments depend on taxing their citizens in order to generate revenue to provide for their citizens' needs, to pay for basic government functions, and to advance their political agendas. However, the relation between a government and its taxpaying citizens is special. Taxpaying citizens expect a return on their "hard earned dollars." They

want a *say* in how *their* money is spent, and they want their government to provide for their basic needs, to protect and promote their interests, and to reflect their will. In other words, there will always be pressure for a government that taxes its citizens to be responsive to them. If it is not, it risks the citizens bucking the government, avoiding their taxes and, if the situation is dire, replacing the government with a more responsive one.

However, in countries where substantial aid resources flow into the government's coffers, those governments can pursue their agendas without taxing their citizens to the same degree that they otherwise would. Correspondingly, citizens may feel less entitled to demand more from their government, as they can't insist on having more of a say in how "their" money is spent, if it isn't actually *their* money that is being spent. Moreover, in a nation where the government is receiving little tax money from its citizens, and where its citizens most pressing needs are being addressed by outside aid groups, the government can always claim (whether truthfully or not!) that it lacks the resources to do more to help its citizens and that it has established relations with external groups to provide for its citizens' needs. So, again, if the citizens' basic needs are unmet, the government can claim that the fault lies with the aid groups from the world's richest countries, not with its own inadequacies. In this way, too, aid efforts can undermine a government's responsiveness to its citizens' needs, interests, and will. It does this, in part, by shifting the responsibility for the countries' needy from the governments to outside groups. As importantly, it does this by upsetting the normal relationship between a government and its taxpayers; in virtue of which taxpaying citizens expect to have a say in their government's direction and priorities, since *they* are paying for them. (There is a similar problem in many resource-rich countries, in the Middle East and elsewhere, where state control of a country's rich resources enables royal families or ruling elites to push their social and political agendas without depending heavily on taxation to fund those agendas. This, in turn, often enables such governments to be unresponsive to the will of their citizens (Wenar, 2016)).

Finally, consider the old adage, "he who pays the piper, calls the tune." This adage suggests that in poor countries where much of a government's income is derived from external groups, rather than internal taxes, the governments of those countries will have strong reason to be responsive to the outside groups, and much less reason to be responsive to their own citizens.

There are two problems with this. First, each aid group will have its own agenda,

and its own view about the best way of fostering its agenda "on behalf of the needy." Unsurprisingly, there will often be a gap between what the *outsiders* would like to accomplish, and *how* they want to accomplish it, and what the needy *themselves* would like done, and how they would like it done. This raises many troubling questions about paternalism, autonomy, and respect for local people, their values, and their ways of life. Unfortunately, I cannot pursue these here, however these questions will be addressed further in my book, *Being Good in a World of Need*. The second worry is that even if the aid groups are accomplishing great good, so that there are good reasons for the government to support their efforts, it remains true that being responsive to the benevolent and paternalistic aims of outside aid groups is not the same as being directly responsive to one's citizens. But it is the *latter* that is the mark of good governance, not the former.

In sum, Deaton believes that good governance is *necessary* for substantial and lasting social and economic progress in the world's poorest countries, and that good governance *requires* a government's being responsive to its people. Unfortunately, however, international aid efforts in many of the world's poorest nations can undermine the responsiveness of those nations' governments to their citizens. According to Deaton, this helps account for the empirical evidence, showing little substantial and lasting social and economic progress in many poor countries that have received great amounts of outside aid. Further, this helps explain Deaton's counterintuitive claim that, despite their best intentions, aid groups and their donors may actually be doing more harm than good.

Deaton sums up his position as follows:

> *Aid and aid-funded projects have undoubtedly done much good; the roads, dams, and clinics exist and would not have existed otherwise. But the negative forces are always present; even in good environments aid compromises institutions, it contaminates local politics, and it undermines democracy. If poverty and underdevelopment are primarily consequences of poor institutions, then by weakening those institutions or stunting their development, large aid flows do exactly the opposite of what they are intended to do. It is hardly surprising then that, in spite of the direct effects of aid that are often positive, the record of aid show no evidence of any overall beneficial effect* (2013, pp. 305-306).

There are many possible responses to Deaton's view. One of the most natural

responses raises some especially important, and troubling, issues. Let me turn to that next.

PART IV. THE PROBLEM OF INDIVIDUAL VERSUS COLLECTIVE RATIONALITY AND MORALITY.

Many are unconvinced by Deaton's worries. They see his critique as supporting *Effective Altruism*. Now in fact, Effective Altruism is a somewhat amorphous philosophical and social movement whose members share a common commitment to using reason and evidence to determine the most efficient morally permissible way of promoting one or more of the following goals: aiding non-human animals, existential risks to sentient life on Earth, promoting the Effective Altruism movement itself, researching the most efficient way of promoting good, and aiding the world's needy. However, in this article, when I refer to *Effective Altruism*, I am referring to that portion of Effective Altruism which is concerned with identifying and supporting as efficiently as possible the international relief and development organizations that most effectively aid those people in the world's poorest countries facing premature death or severely debilitating conditions as a result of poverty, famine, war, tyranny, ignorance, or disease.

In particular, in response to Deaton, many would argue as follows. Given that many people are in great need, and that many others could help them at little cost to themselves, it is crucial to identify and support the most *effective* aid groups. *Obviously*, we shouldn't be supporting aid groups doing more harm than good, but *equally obviously*, it seems, there *must* be *some* aid groups doing more good than harm, and we *should* be supporting the most effective of *those* groups.

Deaton, himself, seems to offer support for this position. He grants that there have been some successful health initiatives—for example, early vaccination programs for smallpox or polio—where the costs associated with those initiatives may have been worth bearing (Deaton 2013, pp. 308-309). Given this, doesn't it make sense to identify other programs where the costs Deaton worries about are worth bearing given the amount of good to be achieved? Why can't Deaton simply support Effective Altruism? Instead of claiming that we shouldn't be supporting international aid groups operating directly in the world's poorest regions, why shouldn't Deaton contend, more modestly, that we must be very careful about *which* aid groups we support, to make sure that they are, indeed, doing more good than harm?

I believe the key to answering these questions lies in an important, and troubling, fact about practical reasoning; namely, that conflicts that can arise between individual and collective rationality and morality. Parfit has referred to such conflicts as *Each-We Dilemmas* (Parfit 1984, Part One). *Each-We Dilemmas* arise when if each of a number of individuals does what is best, individually, by the lights of a given theory, they, collectively, do worse by the lights of that theory. The most famous examples of Each-We Dilemmas are *Prisoners Dilemmas*.[1] The original Prisoners Dilemma, discussed by game theorists, is a two person dilemma, where if *each* of two prisoners does what is genuinely *best for himself*, according to the standard self-interest theory of individual rationality, they, together, will end up serving a large number of years in prison, say twenty years—ten years each!—rather than a much smaller number of years in prison, say, four years—only two years each! What makes the Prisoner's Dilemma paradoxical is that each prisoner is *fully* aware of the predicament they are in, but there is no *individually* rational way of arriving at the outcome where each only spends two years in jail, rather than ten. Here, we have a conflict between the individually rational choice and the collectively rational choice. From the standpoint of what would be *individually* best for *each* of them, it is clear that *each* should act one way. However, from the standpoint of what would be *collectively* best for the *two* of them, *together*, it is clear that *they* should act another way.

Two person Prisoners Dilemmas are rare in the real world. However, *Many-Person Prisoner's Dilemmas* frequently arise (Parfit 1984, Section 23, pp. 56-62). Unfortunately, it is *often* true that if each of a large group does what is best for herself in self-interested terms, they, together, will be much worse off than they would have been if they had instead done what was best for the group as a whole. So, for example: *each* farmer is better off, in self-interested terms, bringing as many crops to market as possible, no matter what the other farmers do—but, *together*, the farmers would be better off if they brought fewer crops to market, since too many will collapse the crop's price; similarly, *each* fisherman would better off, in self-interested terms, harvesting as many fish as possible, no matter what the other fishermen decide to do—but, *together*, the fishermen would be better off if they harvested fewer fish, since harvesting too many will collapse the stocks and undermine their livelihoods; likewise, each taxpayer would be better off avoiding her taxes, whatever anyone else does—but, *together*, taxpayers will be worse off if they don't pay their taxes than if they do, since a large tax base is

1. There is a massive literature on Prisoner's Dilemmas, too massive to cite here. The *Stanford Encyclopedia of Philosophy* contains a nice article with a useful bibliography on the topic, available online.

necessary for the provision of crucial government services and public goods; and so on.

In *Reasons and Persons*, Parfit showed that analogous *Each-We Dilemmas* can arise for deontological moralities (Parfit 1984, Section 36, 95-98). Specifically, Parfit showed that on deontological theories, people can be in the troubling position where if *each* of them does, *individually*, what she *ought morally* to do, they, *together*, will be doing something which, *collectively*, they *ought* not to do. Parfit's result was fascinating and worrisome. In *Rethinking the Good*, I argued that consequentialist theories can face similar worries (Temkin 2012, Section 3.5, pp. 85-95). In particular, I argued that *if* one accepts certain anti-additive-aggregationist principles for comparing certain outcomes—as most people do—then even on consequentialist theories people can be in the troubling position where if *each* of them does, *individually*, what she *ought*, morally, to do, then they, *together*, will be bringing about an outcome which, *collectively*, they *ought* not to bring about.

One such principle, which most people find plausible, is the following:

> The Disperse Additional Burdens View*: In general, if additional burdens are dispersed among different people, it is better for a given total burden to be dispersed among a vastly larger number of people, so that the additional burden any single person has to bear within her life is "relatively small," than for a smaller total burden to fall on just a few, such that their additional burden is substantial* (Temkin 2012, pp. 67-68).

Here is an example. Suppose an aid group could provide farming equipment to a village, which would relieve hunger in that village for 1,000 people for *fifty years*, or they could provide grain to 4,000,000 people, relieving their hunger for a *single week*. In accordance with the Disperse Additional Burdens View, many people hold that the outcome in which 1000 people had their hunger relieved for *fifty years* would be *better* than the outcome in which 4,000,000 people had their hunger relieved for a *week*, even though in the former case there would "only" be 2.6 million weeks of hunger relief. This is because relieving someone's hunger for *fifty years* has a significant impact on her life, while relieving someone's hunger for only one week has relatively little impact on the overall quality of her life.

Not everyone accepts anti-additive-aggregationist principles. Notoriously, they

are rejected by total utilitarians. But consider the following case (from Temkin 2012, see pp. 34-38, and also 42, 259-264, 339, and 484-488):

> *Lollipops for Life*: In outcome A, countless people live very long lives, and they all have *enormously* satisfying lives along every important dimension of human life, along with, more trivially, *lots* of licks of many different lollipops over the course of their lives; unfortunately, however, A also involves one innocent person suffering unbearable agony for eighty straight years, before dying a slow, lonely, torturous death. By contrast, outcome B involves the same countless people living the same enormously satisfying lives, except that they each receive *one* less lick of a lollipop over the course of their very long lives; however, in B, the innocent person would be spared the agony and painful death, and would instead live a full rich life.

Total utilitarians are committed to the view that if only there were *enough* people each enjoying a *tiny* amount of pleasure from the one extra lick of a lollipop, then A would be better than B. Most people, including most consequentialists, reject the total utilitarian's judgment about my Lollipops for Life case. For certain comparisons, at least, they reject total utilitarianism's simple additive-aggregationist approach in favor of the anti-additive-aggregationist approach of principles like the Disperse Additional Burdens View.

For most people, then, the Disperse Additional Burdens View seems deeply compelling. However, it can give rise to consequentialist *Each-We Dilemmas*. To see this, consider the following example:

> *The Reservoir, the Drowning Child, and the Toxic Watch Battery*: Uhuru is walking by a reservoir where a child is drowning. If she pauses to remove her watch before diving in, the child will suffer severe brain damage. If she doesn't remove her watch, its battery will leach toxic chemicals into the reservoir, increasing its pollution level by a *very* small amount. The reservoir is the main source of water for the region's animal life and 1,000,000 people.

What should Uhuru do? Uhuru might plausibly reason as follows. If she removes her watch first, this will *significantly* impact the child. If she doesn't, this may have a *very small* negative impact on each of the many people and animals who depend on

the reservoir for their water. Since there are *so* many sentient beings using the water, we may suppose that the *total* amount of negative effects will be *larger* if she doesn't remove her watch than if she does. Still, the *distribution* of those effects is *very* different. If she removes her watch, *all* of the negative effects will be borne by one child. If she leaves her watch on, the negative effects will be dispersed across a vast number of sentient beings so that each of their lives would be *barely* impacted. Given this, Uhuru might conclude, in accordance with the Disperse Additional Burdens View, that if she wants to produce the best possible outcome, she should dive in immediately and spare the child severe brain damage.

Suppose that Uhuru is right about this. She would then be acting *rightly* in consequentialist terms. Notice, however, that Uhuru might not be the only person facing such a decision. Suppose that 30,000 others were in a similar predicament. No matter what anyone else did, each might act as Uhuru did, and for the same reasons. In so doing, *each* would produce the *best* of her available outcomes, and so be acting *rightly*, as *individuals*, in consequentialist terms. Still, the *cumulative* impact of 30,000 toxic batteries might be very bad. In particular, while the *individual* negative impact on each sentient being from the increased pollution level of a single watch battery might be very small, the *collective* negative impact of 30,000 batteries might be quite significant. Thus, it might well be that, together, the quite significant negative impact on *millions* of sentient beings would be worse than the negative impact of brain damage on 30,000 children. If so, Uhuru and her peers would be facing a consequentialist *Each-We Dilemma*. If, in accordance with the anti-additive-aggregationist reasoning of the Disperse Additional Burdens View, each *individual* does what is best in consequentialist terms, they, *together*, end up producing an outcome which is worse in consequentialist terms.

We can now see why Deaton might grant that *some* aid groups do more good than harm, and yet resist the Effective Altruist's view that we should identify and support those groups. For Deaton, the issue isn't whether aid groups are doing more good than harm at the *individual* level. His concern is with the *collective* impact of such groups. If the preceding is correct, then it could be that *even if each* of us, *individually, only* supports effective aid groups that are doing more good than harm, it could *still* be the case that, *collectively, we* are doing more harm than good.

I believe these considerations help illuminate Deaton's position, as well as most people's reactions to it. Deaton urges us not to support aid groups operating in some of the world's poorest countries, largely on the grounds that doing so weakens the

local governments' responsiveness to their citizens. But most people find this line of reasoning unbelievable. As individuals, each thinks of the great good that *her* particular contribution might do. She might, after all, *save a life*! By contrast, she thinks that the extent to which *her* individual contribution will weaken a government's responsiveness to its citizens will be *ludicrously* small. Thus, the negative impact that *her* contribution will have on each of the country's many citizens will be *so* small as to not even be measurable. Therefore, in accordance with the Disperse Additional Burdens View, her individual contributions will be doing more good than harm, contrary to what Deaton *seems* to be suggesting.

This reasoning is cogent, so far as it goes. But I believe it misses Deaton's point. Deaton isn't taking the ground-level perspective of what each individual donation is, or is not, accomplishing. Deaton is taking the 30,000 foot view of things. He is looking at the net impact of *vast* numbers of individual acts on behalf of the needy. And what he sees, from that perspective, is that the *collective* negative impact of those vast numbers of individual acts is quite substantial. Thus, while I, individually, may have virtually no impact on a government's responsiveness to its citizens; we, together, can have a substantial impact on its responsiveness. And, of course, Deaton believes that, ultimately, a government's responsiveness to its citizens is the crucial component for substantial and lasting social and economic progress.

This is why Deaton urges us not to support aid groups. His contention needn't be that each of us, *individually*, is doing more harm than good. It is, rather, that we, *collectively*, are doing more harm than good. As we have seen, if principles like the Disperse Additional Burdens View are correct, the latter can be true, even if the former is not.

In his famous article, "Famine, Affluence, and Morality," Singer implied that people of good conscience may have to do *more* than they otherwise would to aid the needy, given that not enough other people who are able to help actually do so (Singer 1972, pp. 232-233). Ironically, Deaton's view is almost the opposite. He believes that people of good conscience may have to do *less* than they otherwise would to aid the needy, given that so many other people are doing the same thing! Underlying Deaton's view is the conviction that, collectively, the direct, indirect, and interaction effects of such efforts do more harm than good.

Conflicts between individual and collective rationality and morality are profoundly troubling. Arguably, they lie at the root of many of our most pressing social and political problems, and they can be particularly intractable. Indeed, climate

change, global warming, pollution, destruction or depletion of natural resources, protectionist economic policies, refugee crises, sky rocketing medical and insurance costs, restricted immigration policies, and the proliferation of weapons of mass destruction can all be seen as manifestations of such conflicts among people and/or nations. Unfortunately, the domain of obligations to the needy is no exception to this. Effective Altruists may be right that many effective aid groups are doing more good than harm. Given this, perhaps *each* of us, *individually, ought* morally to support such groups. Yet, despite this, it is possible that, *collectively, we* ought *not* to support such groups since, if we do, we, *together*, may do more harm than good. If that is our real-world predicament, then it may be very clear what each *individual* should do, and also very clear what we, *together*, should do; but what would remain painfully *unclear* is how one could defensibly reconcile the two perspectives.

PART V. A RESPONSE TO DEATON.

In this section, I want to briefly reconsider Deaton's view that we shouldn't provide direct aid to people in the world's poorest countries, because doing so contributes to poor governance within such countries. Importantly, Deaton offers numerous suggestions for how we might try to *indirectly* help people in the world's poorest countries (Deaton 2013, pp. 312-324). In doing this, he quotes favorably the economist Jagdish Bhagwati, who claimed that "it is hard to think of substantial increases in aid being spent effectively *in* Africa. But it is not so hard to think of more aid being spent productively elsewhere *for* Africa" (Deaton 2013, pp. 318-319). However, it is worth noting that many of the concrete suggestions that Deaton offers for how we might help people in some of the world's poorest countries would only help badly-off people at some time in the future, not those whose current desperate plight cries out for immediate amelioration.

Recall the so-called Paradox of Poverty, which holds that aid is unnecessary in countries with good governance, and unhelpful in countries with bad governance. If this is right, then there is *already* poor governance in those countries where so many desperate people need help. So, it isn't as if withholding aid will prevent there from being poor governments in such countries. They are already there, with or without our interventions! Hence, it appears that our choices are between letting needy people suffer, while they are ruled by unresponsive governments; or helping them out, while they are ruled by unresponsive governments! If, in fact, *those* are our choices, it may

seem plain that we ought to do the latter, notwithstanding the ways in which outside aid can undermine a government's responsiveness.

Deaton seems to be suggesting that we should let people suffer now, on the chance that doing so may lead to long-term changes in their government's responsiveness, which, in turn, may eventually lead to substantial long-term social and economic progress. Perhaps Deaton thinks that if poor governments couldn't count on outside resources to fund their agendas, and take care of their needy, they would *have* to adopt policies that would generate tax revenues to enable them to advance their agendas, remain in power, and deal with their countries' problems. Presumably, the most sustainable way to do this would involve adopting policies that would eventually transform their societies' neediest members from being drains on their societies' resources, to being contributors to their societies' tax bases.

Such an approach has some intuitive plausibility. Still, one might think it is a pretty cold-hearted and risky approach—as it abandons the present needy to their cruel fate, with no guarantees that doing so *will* lead to the necessary changes in government responsiveness that Deaton champions. Notice, it *could* turn out that the expected harms of letting many needy suffer now, might be outweighed by the expected benefits of far more people not being needy for decades to come, even if the expected harms are a virtual certainty, while the expected benefits are less likely to be realized than not. In that case, Deaton's somber advice would be endorsed by Effective Altruism. Even so, we might balk at following it.

Consider the standard deontological views that I ought to save my mom, rather than five strangers; or that I ought not to break my promise, to stop five others from breaking their promises; or even that I ought not to break my promise to you today, even if that is the only way of *my* keeping five other promises in the future. Similarly, consider the almost universal appeal of heroic rescues. There is something uplifting, noble, and morally compelling about searching through the rubble days after a major earthquake on the off chance of finding someone still alive, even though the price of doing so would almost never be justified on cost-effectiveness grounds.

These observations remind us there is much more to morality, and to being a good person, than doing the most good that we can. A thoroughly decent person will be virtuous, and will also give weight to deontological considerations at odds with maximizing the good. This is why many of us may feel queasy about Deaton's recommendations, *even if* we accept that they might be supported by long-term, impartial, cost-effectiveness calculations. When we learn of people suffering from the

ravages of war, illness, or natural disasters, many morally relevant factors move us to ease their plight. Perhaps we could do more total good by pursuing other, more cost-effective, long-term goals. However, for many of us, we are not prepared to sacrifice the current needy on the altar of need minimization. We would be fools, or worse, to ignore Deaton's important considerations. However, we must balance those considerations against all of the other considerations relevant to how a decent person responds to the plight of the needy (Temkin 2017a and forthcoming).

PART VI. CONCLUSION.

Some people will be frustrated or even angered by this article. Here I sit, comfortably speculating about various *possible* negative effects that aid groups may produce. In doing this, I provide ammo for all those who selfishly pursue materialistic lifestyles of wasteful consumption, and do nothing to aid the needy. Worse, I haven't offered empirical evidence to support the concerns that I have raised. Meanwhile, millions of flesh and blood innocents are dying or suffering from easily preventable hunger or disease. Don't I know that even *raising* these worries may contribute to needless suffering?

I understand such reactions. Indeed, I have lain awake many nights with the same concerns. My hope is that if my worries can be laid to rest, that will be shown quickly, and if they cannot, people will rethink their assumptions and proceed along a safer, sounder path.

As for gathering and assessing empirical data, I must leave that to the social scientists. My job, as a philosopher, is to help identify both empirical and normative issues relevant to our obligations to the needy, which I have done. Also while I deeply worry that this article may do more harm than good, I also worry that Deaton may be right, and that my previous one-sided approach to thinking about the needy may have been doing more harm than good. There are practical dangers in taking up any complex, morally important topic, but also practical dangers in failing to take up such topics, and letting society's dominant social mores shape people's views about them. The philosopher's job is to carefully and honestly examine such topics, and see where the arguments lead. This article engages in that enterprise, even if only partially.

Unfortunately, I have had to leave so many pertinent empirical and normative questions open for now, that I cannot offer too much concrete advice here. However, let me conclude this article with various considerations to bear in mind, and paths

that still need to be explored, in thinking about how a good person should respond in a world of great need. What I offer, here, are mostly assertions, the arguments for which have been offered elsewhere, or must await another occasion (see Temkin 2004, pp. 349-395 and 409-458).

First, in 1996, the World Health Organization adopted the *DALY—disability-adjusted life year*—as its standard measure for assessing the negative impact of conditions of ill-health (The World Health Organization "Metrics"). Ever since then, many global health experts, and many Effective Altruists, have shared the common approach of measuring the effectiveness of interventions on behalf of the needy in term of the minimization of DALYs. Though understandable, given the importance of health to human wellbeing, our concern for the needy must encompass much more than just health-related goals. Specifically, we must pay attention to deontological-, virtue-, egalitarian-, fairness-, and justice-based reasons for aiding the needy, as well as the consequentialist-based reasons embodied by DALYs.

Second, we must take seriously the fact that to some extent we may be directly or indirectly responsible for the plight of at least some of the world's needy, and this may be true both individually and collectively. This raises a host of complex issues about individual and collective responsibility, and how to trade-off between helping those whose plight we may be partially responsible for, and others whose plight is wholly independent of us, but who may be every bit as needy or more, and whom we may be able to benefit to an even greater extent with equal or fewer resources.

Third, we must face the fact that what each of us, individually, has most reason to do, may be different from what we, together, have most reason to do. As I have argued, tragically, it might be that if *each, individually*, does what she *ought* to do, morally, on behalf of the needy, that *we, collectively*, will *not* be doing what we *ought* to do, morally, on behalf of the needy.

Fourth, there are important moral reasons to personally help the needy, even though this may not do the most good. Similarly, there are moral reasons to not to perform certain jobs, or actions, even though doing so would most benefit the needy.

Fifth, there are moral reasons to focus on *people*, rather than *countries* that are badly off, and it is likely that one will maximize the expected value of one's aid efforts by focusing those efforts in countries with good governance, rather than in countries with poor governance. Since the overwhelming majority of the world's neediest people live in the world's two most populous nations, China and India, it may well be that we should focus more of our efforts to aid the needy in such countries, or even in

richer countries, who have desperately poor inhabitants within their borders, rather than in other desperately poor regions of the world where the problems of poor governance are especially egregious. It is striking, for example, that in 2013, 3.2 million people lived on less than $1.90 a day in the U.S., and another 3.3 million people lived on less than that amount in other high income countries, and also that more people in the U.S are absolutely poor by global standards (5.3 million), than in Sierra Leone (3.2 million) or Nepal (2.5 million), and about the same as in Senegal (5.3 million) (Deaton 2018).

Sixth, notwithstanding the previous point, many moral considerations will support aiding people urgently in need now, even if they live in countries with poor governance, and other available efforts might have greater total expected value. Indeed, in some cases, I believe that we should aid those in dire straits, even if doing so may ultimately do more harm than good.

Seventh, we need social scientists, aid activists, Effective Altruists, and others, to explore even more deeply the probability of any negative effects of aid efforts. In doing this, they must attend to indirect, interaction, long-term, and collective effects, as well as direct, short-term, individual effects. Nothing short of brutal, clear-eyed honesty is acceptable if we hope to answer the critics of international aid and, more importantly, if we really hope to do as much as we can on behalf of the needy.

Eighth, ultimately our aim is to break the cycles of poverty, war, repression, hunger, ignorance, prejudice, and illness that cause people to be needy. Thus, we must guard against aid efforts that indirectly contribute to such cycles by buttressing gangsters, warlords, evil leaders, or repressive or unresponsive governments. We must also identify effective, long-lasting approaches to undermining the root causes of hunger, poverty, and disease. This will need to include not only efforts to improve infrastructure, education, health care, energy production, and such, but efforts to promote equality, justice, human rights, the rule of law, and fundamental changes in the rules that govern national and international social, political, and economic interactions.

Many will dismiss such claims as banal, impractical, and unrealistic. We know how to provide people with mosquito nets, and we can get a general consensus for malaria eradication. But, many will claim, substantial changes in the global political and economic order are another matter, with too many powerful interests lined up against it for it to be feasible. Perhaps. Yet, as the old proverb states, the journey of a thousand miles begins with one step, and if we hope to one day attack the *roots* of

the problem of global need, and not merely its symptoms, then we must map out our journey, and begin taking its crucial first steps, however slow and hopeless they may seem.

The point about attacking the roots of global need, and not merely its symptoms is, of course, a familiar one. Citing another well-known proverb—feed a man a fish, and he eats for a day; teach a man to fish and he eats for a lifetime—aid groups have long trumpeted the importance of *development* efforts, and not merely *relief* efforts (though this long-held staple of many international aid groups in not uncontroversial, and in recent years there has been significant pushback against it (see Ferguson 2015 and Van Parijs 1995)). Still, with a few notable exceptions—such as *GiveDirectly*, which focuses on direct cash transfers to the poor rather than development as the best way of aiding the needy and which has been endorsed by GiveWell as one of the most effective international aid organizations—aid groups have tended to focus on goals like improving water supplies, farming techniques, education, infrastructure, eradicating diseases, and empowering women, goals that seem fairly achievable via outside interventions. In doing this, perhaps aid groups have hoped that necessary social, political, and economic changes would accompany the improvements they achieve. Such hope is not entirely unreasonable, especially with advances in education and female empowerment. Still, I believe we need to try to identify more direct ways of effectively addressing the many systemic factors giving rise to the needy, including the many institutions, rules, and laws that regulate international political and economic relations.

In choosing which aid agencies to support, one will inevitably make trade-offs. One could devote one's resources to relief efforts; to development efforts; or to long-term social, political, and economic changes. A fourth approach would devote different portions of one's resources to each of the three aims. Individually, it may not matter which of these approaches one adopts. I'm not sure about that. However, collectively, I believe that we, together, should adopt the fourth approach. Moreover, on my pluralistic approach, I believe there will be many cases where we ought to aid the needy even though, in terms of pure cost-effectiveness, that money could be better spent elsewhere.

I remain convinced, as I have been throughout my life, that the well-off are open to serious moral criticism if they ignore the plight of the world's needy. Unfortunately,

what one should do in light of that truth is much more complex, and murky, than most people have realized.

Acknowledgements

This article is based on the third of my three 2017 Uehiro Lectures, sponsored by the Oxford Uehiro Centre for Practical Ethics, and delivered at Oxford University in November, 2017. I would like to express my appreciation to the Uehiro Lectures Selection Committee and, especially, Julian Savulescu, for inviting me to give the Lectures; to Rachel Gaminiratne and Miriam Wood for organizing them; and to the Uehiro Foundation on Ethics and Education for its generous support of the Lectures. I would also like to express my deep debt and gratitude to Peter Singer and Angus Deaton, each of whom, in their own ways, compelled me to address these topics; to Derek Parfit, my lifelong teacher, mentor, colleague, and friend; and, most of all, to my parents Lee and Bud Temkin, who instilled in me from an early age a deep concern about the world's needy.

REFERENCES

Banerjee, A. V. *Making Aid Work*, Cambridge: MIT Press (2007).

Deaton, A. *The Great Escape: Health, Wealth, and the Origins of Inequality*, Princeton: Princeton University Press (2013).

——— "The U.S. Can No Longer Hide From Its Deep Poverty Problem," *The New York Times*, (2018) https://www.nytimes.com/2018/01/24/opinion/poverty-united-states.html (accessed on 3 February, 2018).

Easterly, W. *The White Man's Burden: Why the White's Efforts to Aid the Rest Have Done So Much Ill and So Little Good*, Oxford: Oxford University Press (2006).

Ferguson, J. *Give a Man a Fish*, Durham: Duke University Press (2015).

GiveDirectly https://www.givedirectly.org/ (accessed on 3 February, 2018).

GiveWell https://www.givewell.org/charities/top-charities (accessed on 3 February, 2018).

Kuhn, S. "Prisoner's Dilemma," The *Stanford Encyclopedia of Philosophy,* https://plato.stanford.edu/entries/prisoner-dilemma/ (accessed on 3 February, 2018).

Moyo, D. *Dead Aid: Why Aid Is Not Working and How There Is Another Way for Africa*, London: Penguin Books (2010).

Parfit, D. *Reasons and Persons* (Oxford: Oxford University Press (1984).

Singer, P. "Famine, Affluence, and Morality," *Philosophy and Public Affairs*, Vol. 1, No. 3 (1972).

Temkin, L. "Why Should America Care?" *Ag Bioethics* Forum 11, no. 1, pp. 9-15 (June, 1999).

———"Thinking about the Needy, Justice, and International Organizations," *Journal of Ethics*, 8, pp. 349-395 (2004a).

———"Thinking about the Needy: A Reprise," *Journal of Ethics*, 8, pp. 409-458 (2004b).

——— *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*, New York: Oxford University Press (2012).

———2017a Uehiro Lectures, *Obligations to the Needy*, https://www.practicalethics.ox.ac.uk/uehiro-lectures-2017 (accessed on 30 January, 2018).

———2017b Second Uehiro Lecture, audio recording http://media.philosophy.ox.ac.uk/uehiro/MT17_UL_Temkin2.mp3 (accessed on 30 January, 2018).

———*Being Good in a World of Need*, Oxford: Oxford University Press (forthcoming).

Temple, J. "Aid and Conditionality," in the *Handbook of Development Economics*, dedited by Dani Rodrik and Mark Rosenzweig, Amsterdam: Elsevier (2010).

Wenar, L. "Poverty Is No Pond: Challenges for the Affluent" in *Giving Well: The Ethics of Philanthropy*, edited by Patricia Illingworth, Thomas Pogge, and Leif Wenar, New York: Oxford University Press (2011).

———*Blood Oil: Tyrants, Violence, and the Rules That Run the World*, New York: Oxford University Press (2016).

Van Parijs, P. *Real Freedom for All: What (if anything) can justify capitalism?*, Oxford: Clarendon Press (1995).

The World Health Organization, "Metrics: Disability-Adjusted Life Year (DALY)" http://www.who.int/healthinfo/global_burden_disease/metrics_daly/en/ (accessed on 3 February, 2018).

Each-We Dilemmas and Effective Altruism

MATTHEW CLARK

University of St Andrews

THERON PUMMER

University of St Andrews

ABSTRACT

In his interesting and provocative article 'Being Good in a World of Need', Larry Temkin argues for the possibility of a type of Each-We Dilemma in which, if we each produce the most good we can individually, we produce a worse outcome collectively.[1] Such situations would ostensibly be troubling from the standpoint of Effective Altruism, the project of finding out how to do the most good and doing it, subject to not violating side-constraints (MacAskill, forthcoming, p. 5). We here show that Temkin's argument is more controversial than it may appear initially regarding both impartiality and goodness. This is because it is both inconsistent with (i) a plausible conception of impartiality (Anonymity) and inconsistent with (ii) the standard view of goodness (the Internal Aspects View). Moreover, because (i) and (ii) are entailed by the sense of 'impartial goodness' that Effective Altruism tentatively adopts, Temkin's argument is less relevant to Effective Altruism than he suggests.

1. FROM DISPERSE ADDITIONAL BURDENS TO EACH-WE DILEMMAS

Consider the following principle, which Temkin claims most people find 'deeply compelling' (Temkin, 2019, p. 13).

1. With Temkin, we are here concerned with the *impartial* goodness of outcomes (see Temkin, 2019, p. 13).

> Disperse Additional Burdens*: In general, if additional burdens are dispersed among different people, it is better for a given total burden to be dispersed among a vastly larger number of people, so that the additional burden any single person has to bear within her life is 'relatively small', than for a smaller total burden to fall on just a few, such that their additional burden is substantial* (Temkin, 2019, p. 12; Temkin 2012, chapter 3).

Crucially, this principle is intended to make a claim about when one outcome is better, impartially speaking, than another.[2] Next consider Temkin's case.

> Reservoir*: Uhuru is walking by a reservoir where a child is drowning. If she pauses to remove her watch before diving in, the child will suffer severe brain damage. If she doesn't remove her watch, its battery will leach toxic chemicals into the reservoir, increasing its pollution level by a very small amount [slightly burdening each of the 1,000,000 people who drink water from it]* (Temkin, 2019, p. 13).

Suppose that Uhuru is choosing between the child receiving an additional burden of size 10,000 ('substantial') and 1,000,000 people each receiving an additional burden of size 1 ('relatively small'). Independently of whether Uhuru *should* dive in immediately, Disperse Additional Burdens entails that she produces the *impartially best outcome* by doing so.

But now suppose Reservoir is iterated. Suppose there are 30,000 other people, each at the same reservoir as Uhuru, in a situation exactly similar to hers. That is, in front of each of these 30,000 people is a drowning child, and each could either dive in immediately thereby saving the child in front of them but increasing the reservoir's pollution level by a very small amount, or instead remove their watch before diving in, thereby avoiding polluting while saving the child but only after the child has suffered severe brain damage. According to Disperse Additional Burdens, while by diving in immediately *each* of these 30,000 people would produce the best outcome they can individually, they *together* produce an outcome which is worse than had they all removed their watches before diving in. For they together would be bringing about an additional burden of size 30,000 for each of the 1,000,000 people drinking

2. Note that Disperse Additional Burdens might instead be formulated as a deontic principle, concerning the choice-worthiness of actions rather than the betterness of outcomes. Ross has similarly argued, in response to Temkin, that Person-Affecting Views should be understood as deontic rather than axiological. For discussion, see: Ross, 2015; Temkin, 2015.

the reservoir water rather than bringing about an additional burden of size 10,000 for each of the 30,000 drowning children, and we can suppose that according to Disperse Additional Burdens the latter is better than the former.

If Disperse Additional Burdens is true, then there are situations in which, though each of us produces the impartially best outcome we can *individually*, we together produce an impartially worse outcome *collectively* (Temkin, 2019, p. 12). This is one sort of *Each-We Dilemma*.[3]

2. IS DISPERSE ADDITIONAL BURDENS PLAUSIBLE?

We contend that Temkin's reliance on Disperse Additional Burdens as a principle of impartial goodness both makes his argument more controversial than it may appear initially, and less relevant to Effective Altruism than he suggests. First, the initial plausibility of Disperse Additional Burdens is, at least in many cases, best explained by other less controversial principles. Moreover, Disperse Additional Burdens is both inconsistent with (i) a plausible conception of impartiality (Anonymity) and inconsistent with (ii) the standard view of goodness (the Internal Aspects View). As (i) and (ii) are entailed by the sense of 'impartial goodness' that Effective Altruism tentatively adopts, Temkin's argument is less relevant to Effective Altruism than he suggests.

There are several considerations, besides Disperse Additional Burdens, that may be able to account for our intuitions about Reservoir, and its iterated version. For example, our intuitions may be influenced by the view that as Uhuru is in an emergency rescue situation involving a face-to-face encounter with a child, she ought to dive straight in independently of whether she produces the impartially best outcome.[4] Moreover, given realistic assumptions about the low probability of harm and costs of deliberation, the best strategy is to dive in unthinkingly. We are also subject to a mistaken general tendency to ignore small effects on large numbers of people (Parfit, 1984, ch.3).

In addition to difficulties finding positive support for Disperse Additional Burdens, we note that it is inconsistent with a plausible conception of impartiality,

3. Temkin borrows this term from Parfit, 1984, chapter 4 (note that Temkin uses 'Each-We Dilemma' in an arguably broader sense than Parfit).

4. For discussions of the potential moral relevance of distance and salience, see: Unger, 1996; Kamm, 2007; Woollard, 2015; Temkin, *The 2017 Uehiro Lectures*; Chappell, forthcoming; Mogensen, forthcoming.

according to which *who in particular* has a given well-being level makes no difference to which of two outcomes is better (that is, impartially better). On this conception, the outcome in which Ann is at level 1, Beth is at level 10, and Cathy is at level 100 is as good as any outcome in which *someone* is at level 1, *someone* is at level 10, and *someone* is at level 100, regardless of who in particular these people are. Call this conception of impartiality *Anonymity*.[5]

To see that Disperse Additional Burdens is inconsistent with Anonymity, compare the following two outcomes, A and B, each containing the same ten people, P1 through P10:[6]

	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10
A	1	2	3	4	5	6	7	8	9	10
B	3	4	5	6	7	8	9	10	11	2

(Numbers in each row represent well-being levels.)

If Anonymity is correct, then the conclusion that B is better than A cannot plausibly be resisted. The worst-off person in B (P10) is at a higher level than the worst-off person in A (P1), the second worst-off person in B (P1) is at a higher level than the second worst-off person in A (P2), and so on. However, P10 is much worse off in B than in A (losing 8 units of well-being) while each other person loses only 2 units of well-being if A comes about rather than B. Thus, assuming that a loss of 8 units is a 'substantial' burden whereas a loss of 2 units is 'relatively small', Disperse Additional Burdens entails that A is better than B. So, Disperse Additional Burdens is inconsistent with Anonymity.

The foregoing highlights one way in which Disperse Additional Burdens is more controversial than it may have appeared initially. Moreover, Effective Altruism is at least tentatively about promoting well-being *impartially*, that is, *counting everyone's well-being equally*, in a way that respects Anonymity (MacAskill, forthcoming, p. 5).

5. Since Anonymity is about axiology only, it is compatible with 'person-tracking' non-consequentialist principles. For relevant discussion, see: Parfit, 2003; Broome, 2004, p. 135; Otsuka, 2018; and Brown, unpublished.

6. Inspired by Parfit, 2003, footnote 16 'musical chairs'; also see Temkin, 2012, pp. 440-445 'progressive disease'.

Disperse Additional Burdens is also inconsistent with the standard view of goodness, as given by the:

> Internal Aspects View: *For any outcome O, O has a unique degree of goodness, determined solely by O's internal features; and one outcome is better than another if and only if it has a higher degree of goodness* (Temkin, 2012, p. 370).

In the example above, A's degree of goodness, as determined solely by its internal features, is lower than B's degree of goodness, as determined solely by its internal features. Therefore, according to the Internal Aspects View, B is better than A. This is inconsistent with Disperse Additional Burdens.

Over many years, Temkin has argued against the Internal Aspects View, and instead defended the *Essentially Comparative View*, according to which, for at least some outcomes, the degree of goodness of an outcome is relativized to comparisons, i.e. there is no fact about how good an outcome is *on its own*, but only relative to what other outcomes it is compared with (Temkin, 2012, p. 371). But we believe that the Internal Aspects View is the standard view *for good reason*. It is the sort of view that nearly everyone finds very plausible, upon grasping it (Huemer, 2013).

It should accordingly be accepted, absent sufficiently strong countervailing considerations. For reasons we cannot explain here, we doubt anyone has presented such considerations.[7]

In his article Temkin suggests a further line of support for Disperse Additional Burdens. He presents his *Lollipops for Life* case, suggesting most people judge, contrary to total utilitarianism,[8] that an outcome containing a miserable innocent person alongside countlessly many well-off people is worse than an outcome containing a flourishing innocent person alongside the very same countlessly many well-off people, where they each get one fewer lollipop lick. However, standard explanations of such judgements—that no number of lollipop licks is as good as a flourishing life,[9] that the well-being of the least well off is of special importance for the goodness of out-

7. For responses to Temkin's 'spectrum arguments' against the Internal Aspects View, see: Pummer, 2018; and Nebel, 2018.

8. Total utilitarianism holds that one outcome is better than another if and only if it contains more well-being in total.

9. Consider, for example, 'value superiority' views in the tradition of Mill's claims about higher and lower pleasures. For relevant literature, see: Arrhenius and Rabinowicz, 2015; Parfit, 2016; and Clark, unpublished.

comes[10]—are consistent with the Internal Aspects View of goodness. Consequently, we find it somewhat misleading when Temkin writes, '[f]or certain comparisons, at least, [most people] reject total utilitarianism's simple additive-aggregationist approach in favour of the anti-additive-aggregationist approach of principles like the Disperse Additional Burdens View' (Temkin, 2019, p. 13). Even if Lollipops for Life yields strong reason to reject total utilitarianism's simple additive-aggregationist approach, it does not give us compelling reason to accept Disperse Additional Burdens over the Internal Aspects View of goodness.

Finally, Effective Altruists are concerned with quantities like 'number of lives saved', 'number of children dewormed', and so on, as *amounts of good*, which do not depend on what alternatives are available. These amounts of good feature centrally in cost-effectiveness comparisons across different sets of alternatives, and combine with subjective probabilities in expected utility calculations. The Effective Altruist project, so construed, presupposes the Internal Aspects View. At the same time, it does *not* presuppose total utilitarianism, or what Temkin calls an 'additive-aggregationist approach'.

3. EFFECTIVE ALTRUISM AND COORDINATION

Nothing we have argued implies that there are no practical issues of coordination, or of risks of systemic negative side-effects, that are of extreme relevance to Effective Altruism.[11] On the contrary, these and related issues are rightly very much on the minds of Effective Altruists. Especially in recent years, Effective Altruists have been developing solutions to coordination problems, and building mechanisms of effectively 'doing good together' (Dietz, 2018; Collins, forthcoming).

Coordination problems can indeed give rise to Each-We Dilemmas where, if we each produce the best outcome we can individually, we produce a worse outcome collectively. To see this, consider the following familiar example (Gibbard, 1965):

10. Consider the "well-known anti-additive-aggregationist principles of equality and maximin" (Temkin, 2012, p. 70).

11. On risks of harm, see: Wenar, 2011; and Pummer, 2016.

		You	
		Do nothing	Do X
I	Do nothing	Second-best (0)	Bad (-10)
	Do X	Bad (-10)	Best (10)

If we both do nothing, we each produce the best outcome we can *individually* (0), as doing X would produce a worse outcome (-10). But we together produce a worse outcome (0) *collectively* than had we both done X (10). This sort of Each-We Dilemma can be solved by improving coordination in the sense that if we each do the action that produces the best outcome *collectively*, then each of us also produce the best outcome we can *individually*.

Temkin's Each-We Dilemmas cannot be solved in this way. In these situations, the *only* outcomes that can be produced when we each produce the best outcome *individually* are those that are worse *collectively*. But these are precisely the sorts of situations that Regan, Parfit, and others working on coordination problems, have claimed are impossible.[12] Indeed, they are impossible if the Internal Aspects View is true. The Internal Aspects View entails that there is an outcome we could together produce that is *not worse collectively* than any other we could together produce (since, according to this View, each outcome has a unique degree of goodness).[13] Now assume that we together produce an outcome O that is not worse *collectively* than any other we could together produce. If, by acting in some other way, one of us could produce outcome O+ which is better *individually*, then, according to the Internal Aspects View, outcome O+ is also better *collectively* and we could together produce it. But this contradicts the assumption that O is not worse *collectively* than any other outcome we could together produce.

The putatively troubling Each-We Dilemmas implied by Temkin's argument will not arise for projects that promote impartial goodness as construed by the Internal Aspects View. Effective Altruism is such a project. Consequently, Temkin's argument is less relevant to Effective Altruism than he suggests.

12. For discussion, see: Regan, 1980, pp. 54-55; Parfit 1984, especially section 21 and p. 91; Rabinowicz, 1989; and Temkin 2012 chapter 3, footnote 19.

13. Assuming there is a finite number of available outcomes.

In closing, we would like to emphasize that Temkin does not believe it is *wrong* to individually pursue the Effective Altruist project, as we construe it. In personal communication, dated 27 July 2018, he wrote:

> *[O]n my view, though doing that is likely to be both permissible and laudable, relative to most other things one might do, I don't believe that doing so will also be the MOST laudable thing to do. Nor do I believe that a truly good person will, or should so far as possible, always or even generally be motivated by that project. But surely there will sometimes, and perhaps even often, be times when acting in such a way WOULD be precisely the thing that a good person would, and should, do.*

We agree that pursuing the Effective Altruist project, as we construe it, is permissible and laudable. We go further in advocating for its wider adoption, but it is, in any case, important to recognize the scope of the challenge Temkin raises.

Acknowledgements

For helpful comments, we are grateful to Roger Crisp, Tom Douglas, Ben Sachs, and Larry Temkin.

REFERENCES

Arrhenius, Gustaf & Rabinowicz, Wlodek (2015). Value Superiority. In Iwao Hirose & Jonas Olson (eds.), *The Oxford Handbook of Value Theory*. New York: Oxford University Press.

Broome, John (2004). *Weighing Lives*. Oxford: Oxford University Press.

Brown, Campbell (unpublished). Anonymity and Moral Equality.

Chappell, Richard (forthcoming). Overriding Virtue. In Hilary Greaves & Theron Pummer (eds.), *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Clark, Matthew (unpublished). Continuous Superiority.

Collins, Stephanie (forthcoming). Beyond Individualism. In Hilary Greaves & Theron Pummer (eds.), *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Dietz, Alexander (2019). Effective Altruism and Collective Obligations. *Utilitas* 31 (1):106-115.

Gibbard, Allan. (1965). Rule-utilitarianism: Merely an illusory alternative? *Australasian Journal of Philosophy* 43 (2):211-220.

Huemer, Michael (2013). Transitivity, Comparative Value, and the Methods of Ethics. *Ethics* 123 (2):318-345.

Kamm, Frances (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

MacAskill, William (forthcoming). The Definition of Effective Altruism. In Hilary Greaves & Theron Pummer (eds.), *Effective Altruism: Philosophical Issues*. Oxford: Oxford University Press.

Mogensen, Andreas (forthcoming). The Callousness Objection. In Hilary Greaves & Theron Pummer (eds.), Effective Altruism: Philosophical Issues. Oxford: Oxford University Press.

Nebel, Jacob. (2018). The Good, the Bad, and the Transitivity of Better Than. *Noûs* 52 (4):874-899.

Otsuka, Michael (2018). How it makes a moral difference that one is worse off than one could have been. *Politics, Philosophy and Economics* 17 (2):192-215.

Parfit, Derek (1984). Reasons and Persons. Oxford: Oxford University Press.

——— (2003). Justifiability to each person. *Ratio* 16 (4):368–390.

——— (2016). Can We Avoid the Repugnant Conclusion? *Theoria* 82 (2):110-127.

Pummer, Theron (2016). Risky Giving. *The Philosophers' Magazine* 73 (2):62-70.

——— (2018). Spectrum arguments and hypersensitivity. *Philosophical Studies* 175 (7):1729-1744.

Rabinowicz, Wlodek (1989). Act-utilitarian prisoner's dilemmas. *Theoria* 55 (1):1-44.

Regan, Donald. (1980). *Utilitarianism and Co-Operation*. Oxford: Oxford University Press.

Ross, Jacob (2015). Rethinking the Person-Affecting Principle. *Journal of Moral Philosophy* 12 (4):428-461.

Temkin, Larry (2012). *Rethinking the Good: Moral Ideals and the Nature of Practical Reasoning*. New York: Oxford University Press.

——— (2015). Rethinking Rethinking the Good. *Journal of Moral Philosophy* 12 (4):479-538.

——— (2017). *The 2017 Uehiro Lectures*, The Oxford Uehiro Centre for Practical Ethics.

——— (2019). Being Good in a World of Need: Some Empirical Worries and an Uncomfortable Philosophical Possibility. *Journal of Practical Ethics* 7(1): 1-24.

Unger, Peter (1996). *Living High and Letting Die: Our Illusion of Innocence*. New York: Oxford University Press.

Wenar, Leif (2011). Poverty is No Pond: Challenges For the Affluent. In Patricia Illingworth, Thomas Pogge & Leif Wenar (eds.), *Giving Well: The Ethics of Philanthropy*. pp. 104-132. New York: Oxford University Press.

Woollard, Fiona (2015). *Doing and Allowing Harm*. Oxford: Oxford University Press.

Being Good in a World of Uncertainty: A Reply to Temkin

THEODORE M. LECHTERMAN

Centre for Advanced Studies Justitia Amplificata

ABSTRACT

This reply affirms Temkin's critical perspective on effective altruism but seeks to draw out its constructive implications. It first encourages Temkin to defend the practical urgency of global poverty in the face of doubts about aid effectiveness. It then argues for a more holistic conception of effectiveness to mitigate these doubts. It considers some alternative aid strategies that respond to this broader conception. Finally, it exhorts effective altruists to think more seriously about the reform of global institutions.

INTRODUCTION

Temkin's critique of effective altruism stands apart from others in at least three important ways. First, it issues from someone whose commitment to the core tenets of effective altruism is beyond dispute. Second, while other philosophers have fretted about the demands that effective altruism makes on altruistic agents, Temkin helpfully redirects our attention to the effects of the movement on its intended beneficiaries. Third, the account is impressively comprehensive: it illustrates in great detail the challenges that foreign assistance poses to a wide range of values, including underappreciated dimensions of political morality like voice, autonomy, and respect.[1]

1. Here I'm mainly referring to Temkin's forthcoming book, which expands on the ideas published in this journal.

Temkin's is also the first critical perspective that I myself have found thoroughly convincing.

Like Temkin, I believe that global poverty makes considerable demands on those it spares, but what those demands actually amount to in practice is really quite murky (Lechterman forthcoming). Here, I want to explore some potential practical upshots of this conclusion. If one accepts that supporting foreign assistance initiatives leads us into a moral minefield, how can we best avoid, or navigate through, this dangerous terrain?

1. WHY POVERTY?

Some effective altruists might see the challenges of mitigating global poverty as the final nail in the coffin for this "cause area."[2] For many effective altruists, devoting resources to global poverty has become a hard sell even in the absence of Temkinian worries. This is because of the growing awareness of other catastrophes, which strike some as more morally urgent or at least more tractable. Measured against the importance of preventing the extinction of humanity, warding off the annihilation of the Earth, or reducing the horrific mistreatment of nonhuman animals, the misery of a "mere" two billion poor persons looks to some like a "rounding error" (Matthews 2015). Those who pair doubts about the relative significance of global poverty with emerging skepticism about strategies to combat it may be convinced to abandon this problem entirely.

We'd do well to pay attention to other important social problems, particularly ones that have suffered from neglect. But I suspect that Temkin would strongly resist the implication that we should thereby turn our backs on the global poor. And one thing we might ask of Temkin is a clearer justification for why global poverty should remain an urgent priority in light of these challenges. It's not obvious to me that Temkin's pluralist view of practical rationality can successfully supply this justification. For Temkin, acting well requires recognizing and responding to a host of virtue-based, agent-relative, and agent-neutral considerations. One who accepts this picture might still maintain that the agent-neutral reasons to mitigate existential risk are so compelling that they swamp the lingering virtue-based and agent-relative reasons to assist the global poor.

2. Effective altruists often refer to competing objects of beneficence as "cause areas." See, e.g., www.causeprioritization.org

Defending the priority of global poverty in the face of competing causes remains easier for those who view our relationship to the needy through a theory of global justice, which mediates our duties to distant others. Take the fact that we participate in, and help to sustain, a global order that unfairly benefits us. Take also the fact that our prosperity depends in nontrivial ways on past wrongs that cast a long shadow (conquest, colonialism, exploitation, and so on). These facts help to generate stringent, agent-relative duties to the global poor that we don't have to other potential targets of assistance.

2. DEFINING EFFECTIVENESS

Assuming the challenge of defending the priority of global poverty can be met, as I believe it can, let me turn to the question of how one might navigate the obstacles to addressing it.

Temkin argues that global poverty relief can raise each-we dilemmas, where what each of us has most reason to do (support the most demonstrably effective aid agencies) conflicts with what we all together have most reason to do (support the long-term reduction of poverty rates). I agree with Temkin that conflicts between individual and collective responsibility are key to understanding this problem (and many others). The arguments that aid initiatives undermine development are powerful. And even if these arguments prove faulty, I believe that reasons to help people directly will conflict with justice-based reasons to support the development of decent and stable institutions. Nonetheless, I also wonder whether looking at this problem from another angle might help us find ways around it.

As Temkin points out, most effective aid agencies either provide humanitarian relief or engage in discrete development projects without accounting for how these efforts will affect a country's institutions, especially over the long term. But perhaps the real problem here lies in the definition of effectiveness. Temkin appears to assume a particular definition of effectiveness that has been common among effective altruists. This definition of effectiveness identifies it with measured impact. To be an effective organization, according to this way of thinking, one must be able to demonstrate sizeable impact on some measurable social indicator. The most reliable way to measure impact is through randomized controlled trials, and this explains why RCTs are a coveted source of information for rating agencies like GiveWell. But, as social scientists continually warn, randomized controlled trials are a limited tool (Clough

2015). They can only track certain kinds of interventions, and they can't measure systemic or long-term effects. Organizations that score highly against these limited criteria may be counterproductive in the long run.

Consider an analogy. From a certain point of view, a diet of simple sugars looks remarkably effective. It appears to ramp up energy and trigger sensations of satisfaction. And if the tests we use can only measure local, short-term effects, sugar will look like the most effective form of nutrition around. But, as we all know, sugar's effects are ultimately short-lived, and a diet high in sugar will destroy the body over the long term. Using the logic of practical dilemmas, we could conclude from this that our reason to consume an effective diet in sugar conflicts with other reasons we have to protect our health. But it would be more natural to say that a sugar-rich diet simply isn't an effective form of nutrition in the first place.

The lesson here is that we can mitigate the apparent each-we dilemma in poverty relief by broadening the criteria we use to define and assess effectiveness. We shouldn't consider an organization effective simply by virtue of its demonstrated ability to improve QALYs (or reduce DALYs). Rather, we should assess organizations on an array of criteria that also track relationships to broader development goals. Were we to do this, we'd probably encounter trade-offs between progress on different dimensions of assessment. Sometimes, we might judge that an organization with outstanding effects on QALYs deserves support despite concerns about its long-term impact. Other times, we might judge that the predicted improvements in QALYs aren't worth the risks that a given intervention poses to institutional development. But we can't make these kinds of comparisons without a richer understanding of effectiveness. And because existing data and methods make it difficult to measure and compare systemic effects, what one has most reason to do at this very moment may be to support further research into making these kinds of holistic evaluations.

3. POWER, AUTONOMY, AND RESPECT

A related problem that Temkin exposes is the undue power that comes along with foreign assistance.[3] Even when they are well-intentioned and sensitively administered, funds from abroad create various pressures that tend to distort community priorities and behavior. Here the problem isn't so much that philanthropy from

3. This, too, is something Temkin addresses more elaborately in the larger project from which this article draws.

abroad will interfere with long-term institutional progress (though it may). Rather, it's that resources from abroad can create or reinforce objectionable social relationships. The existence of large sums of money creates incentives for individuals and communities to abandon careers, lifestyles, and policies that they may have preferred to pursue. To access or maintain the flow of benefits, receiving communities face pressure to ingratiate themselves and genuflect to the whims of donors and their agents. And interactions with well-heeled donors and staff members make receiving communities vividly and bitterly aware of their disadvantages. Feelings of powerlessness, humiliation, and disrespect seem not to resonate strongly with many members of the effective altruist movement. But I think it's important to recognize that these experiences can be just as, if not more, disabling to people than poor physical health (Deveaux 2015).

These kinds of considerations have led some to support unconditional cash transfer programs. Because unconditional cash transfers limit the ability of donors to control and intrude, they appear to minimize relationships of domination and subordination. Cash transfers have much to recommend them. But they are no panacea. Aspects of domination and subordination can easily creep in when there are decisions to make about who receives transfers. Cash transfers can't solve pressing problems that really do require technical expertise and specialized equipment, as in medicine and infrastructure. And foreign-funded cash transfer programs pose a clear threat to political participation and government accountability. If the needy can rely on funds from abroad, they have fewer incentives to make demands on the state.

So, it may be worth considering some additional ways of rendering assistance that acknowledge the values of autonomy and self-respect. Certain kinds of participatory organizational structures that give beneficiaries a voice in decision-making are one option (Krasner and Weinstein 2014). Another option is to treat foreign assistance projects as temporary demonstrations that will be handed over to local control after a given period of time (Reich 2016). A third possibility is to provide support for community organizing, which involves helping communities to identify shared interests and overcome collective action problems on their own (Stout 2010). Community organizing is attractive because it involves minimal outside interference, it aims for maximal inclusion, and it can foster the civic virtues needed for effective political participation.

4. REFORMING THE GLOBAL ORDER

As I suggested above, the global order bears a great deal of causal responsibility for the persistence of severe poverty. Even if one rejects the deontological arguments for reforming the global order, one must still recognize that the most sweeping improvements in poverty rates would most likely come from changes to international rules. Consider international laws that grant dictators property rights in natural resources (Wenar 2015), intellectual property rules that prevent poor countries from producing essential medicines (Pogge 2009), migration policies that limit the movement of labor, or trade policies that disadvantage farmers and fledgling industries in the global South (Risse 2012). These are just a few of the ways in which rich countries collectively exploit poor ones, with dire consequences. These are also massive, complicated, and controversial problems, which may help to explain why effective altruists, in search of opportunities for concrete progress, have been drawn to alternative paths. But suppose that effective altruists were to coordinate and consolidate their efforts for a few years to focus on a single aspect of the global order, one where reform might be feasible. As the movement grows in its size and in its collective wisdom, so does its potential to catalyze institutional change.

Reforming international institutions has its fair share of drawbacks. Chief among these is that the benefits it would generate would likely materialize only in the future, to the neglect of those who are suffering now (Cordelli 2016). And one might believe that the prospects of generating the collective will necessary to influence international rules are simply too dim to be worth pursuing. But it remains striking that effective altruists are so quick to embrace other complex global challenges like the risks from artificial intelligence and asteroid collisions, and so leery of thinking boldly about poverty.

REFERENCES

Clough, Emily. 2015. "Effective Altruism's Political Blind Spot." *Boston Review*, July 14.

Cordelli, Chiara. 2016. "Reparative Justice and the Moral Limits of Discretionary Philanthropy." In *Philanthropy in Democratic Societies,* ed. Rob Reich, Lucy Bernholz, and Chiara Cordelli, 244–67. Chicago: Chicago University Press.

Deveaux, Monique. 2015. "The Global Poor as Agents of Justice." *Journal of Moral Philosophy* 12, no. 2, 125-150.

Krasner, Stephen D., and Jeremy M. Weinstein. 2014. "Improving Governance from the Outside In." *Annual Review of Political Science* 17, 123–45.

Lechterman, Theodore M. Forthcoming. "The Effective Altruist's Political Problem." *Polity*.

Matthews, Dylan. 2015. "I Spent a Weekend at Google Talking with Nerds about Charity. I Came away...Worried." *Vox*, August 10.

Pogge, Thomas. 2009. "The Health Impact Fund: Boosting Pharmaceutical Innovation without Obstructing Free Access." *Cambridge Quarterly of Healthcare Ethics* 18, 78–86.

Reich, Rob. 2016. "Repugnant to the Whole Idea of Democracy? On the Role of Foundations in Democratic Societies." *PS: Political Science & Politics* 49, 466–72.

Risse, Matthias. 2012. *On Global Justice*. Princeton: Princeton University Press.

Stout, Jeffrey. 2010. *Blessed Are the Organized*. Princeton: Princeton University Press.

Wenar, Leif. 2015. *Blood Oil*. Oxford: Oxford University Press.

Medical Crowdfunding, Political Marginalization, and Government Responsiveness: A Reply to Larry Temkin

ALIDA LIBERMAN

Southern Methodist University

ABSTRACT

Larry Temkin draws on the work of Angus Deaton to argue that countries with poor governance sometimes rely on charitable giving and foreign aid in ways that enable them to avoid relying on their own citizens; this can cause them to be unresponsive to their citizens' needs and thus prevent the long-term alleviation of poverty and other social problems. I argue that the implications of this "lack of government responsiveness argument" (or LOGRA) are both broader and narrower than they might first appear. I explore how LOGRA applies more broadly to certain types of charitable giving in developed countries, with a focus on medical crowdfunding. I then highlight how LOGRA does not apply to charitable giving aimed at alleviating the suffering of the absolutely politically marginalized, or those especially vulnerable people to whom governments are never responsive.

1. POVERTY ALLEVIATION AND THE LACK OF GOVERNMENT RESPONSIVENESS ARGUMENT

In his challenging and important paper in this volume, Larry Temkin engages with economist Angus Deaton's argument that foreign aid and other charitable giving to the neediest countries in the world unintentionally does more harm than

good (Deaton 2013, Ch. 7.). We can reconstruct one of Deaton's arguments for this claim as follows:

1. Extreme poverty and other preventable suffering in a given country cannot be effectively alleviated in the long-term unless that country's government is fundamentally responsive to its citizens and their needs (e.g., for healthcare, education, infrastructure, rule of law, etc.).
2. Governments are responsive to their citizens only to the extent that they depend on citizens' support through taxes, votes, and the like.
3. Large amounts of aid (whether direct foreign aid from governments or charitable giving) enable a government to remain in power and attain its goals without citizen support; this undermines the government's reliance on its citizens.
4. Therefore, large amounts of foreign aid make governments unresponsive to citizens.
5. Therefore, large amounts of foreign aid prevent long-term alleviation of extreme poverty and other preventable suffering.

Call this the *Lack of Government Responsiveness Argument*, or LOGRA. LOGRA depends on several controversial empirical claims: (1) that responsive governments are necessary for long-term poverty alleviation, (2) that governments will not be responsive to citizens unless they rely on them for taxes, etc., and (3) that foreign aid enables governments to avoid relying on citizens in these ways. For the sake of argument, assume that these claims are true in at least some cases.

Temkin points out that LOGRA may apply even to demonstrably effective giving of the sort endorsed by Effective Altruism (EA). Givers face what (following Derek Parfit) Temkin calls an *Each-We Dilemma*. These dilemmas are cases in which an individual can bring about the best consequences by the lights of a certain theory by doing one thing, but the overall consequences will be very bad by those same lights if others also do that same thing. An individual donation to an effective aid organization has a massive positive impact on those who are helped and a tiny negative impact on government responsiveness. If we assume that it is better for burdens to be dispersed among many who each bear only a small cost than it is for burdens to be carried by a few who each bear a large cost, then what we ought to do individually is give (so that everyone bears the miniscule burden of a government made slightly less responsive) rather than refrain from giving (so those few who would otherwise be helped bear huge burdens). But LOGRA implies that many donations to effective

aid organizations together risk undermining government responsiveness in a way that inhibits long-term poverty alleviation, which is a massive burden on everyone. Collectively, then, what we all ought to do is refrain from giving.

I will argue that if Temkin is right about this, LOGRA has important implications that go beyond what he addresses in his paper. First, I explore how a version of LOGRA might apply to some forms of domestic charitable giving even in high-income countries with generally well-functioning governments. This means that Each-We Dilemmas concerning charitable giving may be more widespread than Temkin suggests. Second, I suggest that there are groups of especially marginalized people to whom LOGRA does not apply, because governments never rely on them for financial or political support, and therefore have no incentive to be responsive to them even in the best of circumstances. It follows that giving that is narrowly aimed at alleviating the suffering of these groups is not subject to the particular Each-We Dilemma raised by LOGRA.

2. EACH-WE DILEMMAS FOR LOCAL GIVING IN AFFLUENT NATIONS

LOGRA highlights one way in which foreign aid undermines government responsiveness. I am concerned that other forms of charitable giving can similarly undermine government responsiveness. In both affluent and poverty-stricken countries, private individuals or organizations routinely fill gaps in the provision of essential goods and services that can be effectively provided to all in the long term only through state intervention. Although this meets the needs of some individuals in the short term, it risks seriously undermining the government's ability or motivation to meet all of its people's needs in the long term. While there are multiple examples of this, I will focus primarily on how crowdfunding for medical expenses can undermine the political will to fix a broken healthcare system. My arguments are conditional, as they depend on controversial empirical assumptions that I am unable to defend here. However, even if these particular assumptions are false, a similarly structured argument should apply in a wide range of analogous situations.

Crowdfunding typically uses web platforms (such as GoFundMe) to solicit direct donations from friends and strangers. Crowdfunding to cover healthcare expenses is widespread and growing, covering everything from cancer treatment to emergency care to experimental treatments to routine expenses for chronic illnesses. There are

a number of serious ethical problems with medical crowdfunding. Among other concerns, crowdfunding seems to disproportionately and unfairly benefit those who are tech savvy, have wide social networks, are seen as deserving of help, and whose stories are media friendly; ineffectively distributes aid on the basis of sympathy and luck rather than need; and forces recipients to publicly disclose sensitive health information that they might rather keep private in order to receive funding (see Snyder 2016 and Berliner and Kenworthy 2017 for more on these and other criticisms).

Another major worry is that crowdfunding enables governments to shirk their duties. Campaigns are more frequent in areas with less robust health insurance; for example, a randomized survey of GoFundMe crowdfunding campaigns found that "a much larger proportion of campaigns than expected were based in states that chose not to adopt the Medicaid expansion under the ACA" (Berliner and Kenworthy 2017: 237). Jeremy Snyder points out that "the sites allow individuals to address their need for medical care without addressing the underlying causes of these unmet needs" (Snyder 2016: 39). This is exacerbated by the fact that crowdfunding campaigns routinely ignore structural injustice in their pleas for help. For example, a survey of Canadian campaigns found that they focused on the recipient's personal relationships, needs, and altruistic characteristics and "almost universally did not appeal to the perceived injustice of having to resort to crowdfunding by Canadians with an existing entitlement to essential medical care, supporting the concern that medical crowdfunding can obscure systemic injustices" (Snyder et al. 2017 p.367).

If we assume for the sake of argument that the following (admittedly controversial) premises are true, we can generate a LOGRA for medical crowdfunding in affluent countries:

1. Healthcare needs can effectively be met in the long-term only through comprehensive government provision of services.
2. Governments will provide comprehensive healthcare services to all only if politically pressured by their citizens to do so.
3. Citizens will politically pressure governments to act only if they perceive a pressing need.
4. Medical crowdfunding undermines citizens' perceived need to advocate for comprehensive government provision of healthcare.[1]

1. If crowdfunding primarily benefits those who are most skilled at advocating for themselves, it risks undermining the perceived need to advocate for government provision of services among those who are best situated to do this sort of advocacy in particular.

It follows that donors in affluent countries who crowdfund the medical expenses of their friends and neighbors risk undermining the responsiveness of their government in a way that collectively leads to much worse results:

5. Therefore, medical crowdfunding prevents citizens from politically pressuring their governments to provide comprehensive healthcare coverage.
6. Therefore, medical crowdfunding removes incentives for the government to provide comprehensive coverage, which prevents the meeting of long-term healthcare needs.

Snyder rightly notes that "the contribution of any one campaign to these problems is minimal, creating a strong argument that the gain to each user offsets the systemic effects of medical crowdfunding" (Snyder 2016 p.41). He goes on to suggest that "as a result, it is difficult to make the argument that those seeking access to essential medical services through crowdfunding ought not to do so" (*ibid*).

But this argument moves too quickly, because individual donors to medical crowdfunding campaigns potentially face Each-We Dilemmas. Individually, donating to a campaign clearly does good (although surely not the most good you can do with your money in EA terms). Collectively, though, donating to medical crowdfunding campaigns risks undermining the only sustainable long-term solution to meeting everyone's healthcare needs. This is not to say that we should ignore the dire appeals of our family, friends, or strangers for help with their healthcare. With Temkin, I am not ready to "sacrifice the current needy on the altar of need minimization" (Temkin 2019). But we cannot ignore the fact that medical crowdfunding might lead to counterproductive negative effects. We must consider whether our individually good actions are leading to a collectively terrible result, and continue to grapple with the ethics of Each-We Dilemmas in determining whether and how to give to medical crowdfunding campaigns.

Even if medical crowdfunding does not in fact undermine government responsiveness in the area of healthcare, it is worth exploring whether Each-We Dilemmas of this structure arise for other kinds of giving that risk undermining government responsiveness narrowly in other areas (such as funding for scientific research or the arts). For example, consider U.S. billionaires who make major donations to support K-12 public education, such as Mark Zuckerberg's $100 million gift to Newark, N.J.

public schools,[2] or the Gates Foundation's support of public education aimed at improving outcomes for Black, Latino/a, and low-income students.[3] For the sake of argument, assume the (admittedly controversial) claim that the only long-term solution for improving educational outcomes across the board for underrepresented students is state intervention (such as divorcing public school funding from property taxes and providing increased and equitable funding across geographic regions through redistributive taxation). Assume also the (again controversial) claim that well-publicized support of equitable public education initiatives by a handful of billionaires dramatically lessens Americans' perception of the need for different tax policies, and that this makes the U.S. government less responsive to educational inequality. If these claims are true, these billionaire philanthropists face Each-We Dilemmas, and risk undermining the only feasible long-term solution to educational inequality.

The worry that private giving might undermine the political will to solve entrenched social problems is not new. For example, J. A. Hobson wrote in 1914 that

> *Every act of charity, applied to heal suffering arising from defective arrangements of society, serves to weaken the personal springs of social reform... by the softening influence it exercises on the hearts and heads of those who witness it. It substitutes the idea and the desire of individual reform for those of social reform, and so weakens the capacity for collective self-help in society* (Hobson 1914 p.296).

This echoes socialist and leftist critiques of EA which suggest that it ignores institutional factors and "lets capitalism off the hook" with its tendency to "obscure that the ordinary workings of capitalist markets create and exacerbate poverty" (Gomberg 2002 p.55).[4] LOGRA points out a different way in which charitable giving risks undermining institutional effectiveness: not by supporting an exploitative capitalist system, but by preventing capitalist governments from functioning as well as they could.

2. The impact of Zuckerberg's gift has been controversial; see https://www.wsj.com/articles/newarks-100-million-education-debate-1441752228 [Accessed 22/5/19]. Thanks to Alex Dietz for suggesting this example.

3. See http://k12education.gatesfoundation.org/what-we-do/ [Accessed 22/5/19]

4. For further discussion of institutional critiques of EA (and an argument that the only plausible versions of these critiques are consistent with EA principles), see Berkey 2018.

3. LOGRA AND ABSOLUTE POLITICAL MARGINALIZATION

In making his case for LOGRA, Deaton writes, "the need to raise funds exists everywhere, and it will often constrain the ruler to pay attention to the demands of *at least some* of the population" (Deaton 2013 p.295, [my emphasis]). However, neither Deaton nor Temkin pays sufficiently close attention to the fact that governments are responsive *only to some* of the people residing in their countries. Every government is most responsive to certain constituents (e.g., their wealthy political donors). But there are some groups to whom governments are not responsive at all. Call them the *absolutely politically marginalized*. Because they are ineligible to vote and/or do not contribute tax dollars or other material support to the state, governments that are focused on self-preservation will have no *direct* incentive to be responsive to them in any circumstances, even without any financial bolstering from foreign aid. They may have *indirect* incentives if foreign allies put politically pressure on them, or if politically powerful people mobilize on their behalf. But these indirect incentives will likely not lead to the same degree of responsiveness as would direct incentives.

Different groups of people are absolutely politically marginalized in different societies, including (but not limited to): undocumented immigrants, refugees, and those denied citizenship (e.g., the Rohingya in Burma); felons in jurisdictions with felony disenfranchisement; people who do not pay taxes and systematically lack political power (e.g., the chronically homeless and unemployed); and people who are enslaved. Non-human animals are absolutely politically marginalized in an even more extreme way: they have few to no legal rights, and are incapable of directly giving financial or political support to the government. Even if LOGRA succeeds in establishing that foreign aid has serious negative consequences for a country as a whole, it does not follow that aid that is narrowly aimed at alleviating the suffering of the absolutely politically marginalized has the same negative consequences for those marginalized people. For we must consider what would have happened had the aid not been given. Most people will be worse off if an otherwise responsive government becomes unresponsive as a result of aid. But the absolutely politically marginalized cannot be made worse off in this way, since the government is not otherwise responsive to them. And so even large amounts of narrowly targeted aid will not harm them, which means that donors who provide such aid do not seem to face Each-We Dilemmas.

Temkin suggests that one way to avoid LOGRA is focusing aid not on low-income countries with poor governance, but on poor people in middle-income coun-

tries with decent governance, such as China and India. We should also consider focusing our aid efforts on supporting the absolutely politically marginalized in any country. And since it is likely that they will remain in the margins unless their governments become more responsive to them, we must also think more carefully about the value of engaging in political action to encourage governments to become more responsive to the absolutely marginalized. This usually happens only insofar as voters and taxpayers advocate on their behalf: undocumented immigrants protesting their own poor treatment will not motivate a self-interested government to change, but politically powerful people protesting this same poor treatment might. Non-human animals are incapable of advocating for themselves, but the activism of their human supporters has led to major gains in animal welfare laws.

However, we must be cautious that private aid to absolutely marginalized groups does not unintentionally undermine the political will to pressure the government to become responsive to these groups. For if aid groups step in where governments fail in a way that prevents voters from perceiving the dire needs of the absolutely politically marginalized and pressuring their governments in light of this, we risk another version of LOGRA, in which giving to the absolutely marginalized ensures their ongoing marginalization. Ultimately, the implications of LOGRA are potentially both broader (ruling out medical crowdfunding and perhaps other forms of charitable giving in the developed world) and narrower (ruling in giving to the absolutely politically marginalized, unless this itself makes governments less responsive) than it may first appear.

Acknowledgements

Thanks to Brian Berkey, Josh Crabill, Alex Dietz, and Jeremy Snyder for helpful feedback on this paper.

REFERENCES

Berkey, B. (2018) "The Institutional Critique of Effective Altruism." *Utilitas* 30(2): 143-171.

Berliner, L. and Kenworthy, N. (2017) "Producing a Worthy Illness: Personal Crowdfunding Amidst Financial Crisis." *Social Science and Medicine* 187: 233–243.

Deaton, A. (2013) *The Great Escape: Health, Wealth, and the Origins of Inequality*. Princeton: Princeton University Press.

Gomberg, P. (2002) "The Fallacy of Philanthropy." *Canadian Journal of Philosophy* 32(1): 29-66.

Hobson, J. A. (1914) *Work and Wealth: A Human Valuation*. New York: MacMillan.

Snyder, J. (2016) "Crowdfunding For Medical Care: Ethical Issues in an Emerging Health Care Funding Practice." *Hastings Center Report* 46(6): 36-42.

Snyder, J., V. Crooks, A. Mathers, and P. Chow-White. (2017) "Appealing to the Crowd: Ethical Justifications in Canadian Medical Crowdfunding Campaigns." *Journal of Medical Ethics* 43: 364-367.

Temkin, L. (2019) "Being Good in a World of Need: Some Empirical Worries and an Uncomfortable Philosophical Possibility." *Journal of Practical Ethics*. 7:1

Aid Scepticism and Effective Altruism

WILLIAM MACASKILL

University of Oxford

ABSTRACT

In the article, 'Being Good in a World of Need: Some Empirical Worries and an Uncomfortable Philosophical Possibility,' Larry Temkin presents some concerns about the possible impact of international aid on the poorest people in the world, suggesting that the nature of the duties of beneficence of the global rich to the global poor are much more murky than some people have made out.

In this article, I'll respond to Temkin from the perspective of effective altruism—one of the targets he attacks. I'll argue that Temkin's critique has little empirical justification, given the conclusions he wants to reach, and is therefore impotent.

SECTION I

Let us begin by discussing the empirical evidence on aid and economic growth. The majority of Temkin's article discusses the possible negative impacts that foreign aid could have. Towards the end of this article, however, Temkin makes the following striking comment: "Here I sit, comfortably speculating about various *possible* negative effects that aid groups may produce.... I haven't offered empirical evidence to support the concerns that I have raised." [emphasis Temkin's].

This is a surprising approach, to say the least. One might think that, when discussing the ethics of attempting to benefit the global poor—and, especially, when focusing specifically on whether such attempts are effective or not—the vast empirical literature in development economics would be relevant. But Temkin doesn't discuss this literature, instead preferring to rely almost wholly for his empirical claims on a single chapter by a single author of a single book that was intended for a general audi-

ence.[1] There is no doubt that Angus Deaton is an outstanding economist. But is he so good as to warrant ignoring the thousands of other economists who have dedicated their careers to understanding the effects of aid on those in poor countries?

Temkin makes the following methodological remark in defence of his approach: "As for gathering and assessing empirical data, I must leave that to the social scientists. My job, as a philosopher, is to help identify both empirical and normative issues relevant to our obligations to the needy, which I have done." Now, it is true that Temkin's preferred grammatical construction is modal: the words 'possible', 'may', 'might', 'can' and their variants collectively appear over 150 times in his article. And, of course, as philosophers we often should reflect upon the merely hypothetical, because of the broader lessons that we can learn from hypothetical cases.

But I find it hard to believe that this is really the spirit in which Temkin's article is written. His focus is firmly on the very practical question of what we, in the actual world, are obliged to do. And his final sentence is non-modal: "Unfortunately, what one should do in light of [the fact that the well-off who ignore the badly-off are open to serious moral criticism] is much more complex, and murky, than most people have realized."

Perhaps, instead, it's the mere *possibility* of doing harm that means we should potentially refrain from donating to development charities? This would make Temkin's reliance on a single source more justifiable. But it would result in a very weak argument, because there are possible harms from almost every action that we undertake.

For example, almost all the same concerns that Temkin raises for donations would also apply to consumption of goods produced in the developing world by international companies. Perhaps foreign private investment will result in a brain drain away from other industries or from the country itself (if the employees go to work at some other branch of a multinational company). Perhaps those foreign employers will undermine the local governments in order to get their way. Perhaps, even, some of the money that you'd spend on goods produced in the developing world would be used by companies to bribe government officials in order to get a tax break. Is Temkin therefore recommending that we only buy local, too?

In general, laying out what-ifs doesn't get us very far in the ethics of development. All activities we engage in have some chance of harming others, and some chance of benefiting others. What matters is how great the possible benefits are, how great the possible harms are, and the probabilities of each. Temkin, unfortunately,

1. Deaton (2013, ch.7).

doesn't attempt to address these issues,[2] and so, from his article alone, we are left none the wiser about whether development efforts are on average good or bad.

But what should we think about the impact of aid on economic development? It seems to me that there are two reasonable positions that one could have, depending on how sceptical one is towards econometric analysis in this area. If one is sympathetic to the value of econometric methods, then the most natural conclusion is that aid has a positive but modest effect on growth. This view is represented by Jonathan Glennie and Andy Sumner, in a Center for Global Development policy paper:

> *the assertion that aid generally contributes to economic growth, while not proved beyond doubt, is now less contentious in the academic literature than is currently recognised in public policy debate. That is not to say that there is an absolute consensus, nor that there are not important unresolved questions that would need addressing to claim unequivocal proof, but that aid's critics are currently in the academic minority* (Glennie and Sumner 2014).

If one places less credence in econometric methods in this context, then one might reasonably be agnostic either way about aid's effects on growth. The sample size of 80 countries is small, and there are major confounds that are difficult to control for. (If receiving more aid is correlated with lower GDP, is that because aid hinders growth, or because richer governments will give more aid to countries that are poorer?) This view is represented by Owen Barder (himself citing David Roodman) in a report to the British House of Lords: "Given the modest volumes of aid, we should not expect an impact on growth which is bright enough to shine through the statistical fog." (Barder 2011).

SECTION II

However, even if one thinks that there is a lack of robust evidence regarding the effect of bilateral aid on economic growth, that should not lead us to a position of agnosticism about whether one can meaningfully improve the lives of the extreme poor. Crucially, the general debate about aid's impact on economic growth has very

2. See Ravallion (2014) for a discussion of this literature in relation to *The Great Escape*. For a selection of the recent literature on aid, see Burnside and Dollar (2000), Rajan and Subramanian (2008), Clemens et al (2012), and Mekasha and Tark (2013).

limited bearing on the sorts of donations that we should be talking about, such as donations to Against Malaria Foundation, one of GiveWell's top-recommended charities. This is for three main reasons.

First, the debate on aid effectiveness (including the work of Angus Deaton and other aid critics like William Easterly and Dambisa Moyo) is focused on bilateral aid, *not* on non-governmental organisations. But the most obvious donation targets for individuals in rich countries are NGOs.

Second, the vast majority of aid scepticism is aid aimed at economic development, rather than at global health. The track records of these two projects are very different. Though the track record of attempts to foster economic growth is arguably unclear, the track record of global health is astonishing. The eradication of smallpox has saved over 60 million lives since 1980 (more lives saved than if we'd achieved world peace in the same time period) and ⅓ of the funding of the eradication effort came from international aid.[3] Globally, rates of death from measles, malaria and diarrhea are down by 70%.[4] Indeed, even those regarded as aid sceptics are very positive about global health.[5] Here's a quote from Angus Deaton, from the same book that Temkin relies so heavily on:

> *Health campaigns, known as "vertical health programs," have been effective in saving millions of lives. Other vertical initiatives include the successful campaign to eliminate smallpox throughout the world; the campaign against river blindness jointly mounted by the World Bank, the Carter Center, WHO, and Merck; and the ongoing— but as yet incomplete— attempt to eliminate polio* (Deaton 2013 p.104-5).

Later in the book he states: "There may ... be cases in which aid is doing good, at least on balance. I have already made that case for aid directed toward health."(Ibid p.318) Similarly, here's Bill Easterly, another aid sceptic:

> *There are well known and striking donor success stories, like the elimination of smallpox, the near-eradication of river blindness and Guinea worm, the spread of*

3. For further explanation of this, see MacAskill (2015, ch.3).

4. Global IDEA Scientific Advisory Committee (2004). For further case studies in global health, see http://millionssaved.cgdev.org/ [Accessed 22/5/19]

5. These points are made at more length at https://blog.givewell.org/2015/11/06/the-lack-of-controversy-over-well-targeted-aid/ [Accessed 22/5/19]

> *oral rehydration therapy for treating infant diarrheal diseases, DDT campaigns against malarial mosquitoes (although later halted for environmental reasons), and the success of WHO vaccination programmes against measles and other childhood diseases* (Easterly 2009).

In the same post, he summarises his view by commenting that, "even those of us labeled as "aid critics" do not believe aid has been a universal failure. If we give you aid agencies grief on failures, it is because we have seen some successes, and we would like to see more!"

Finally, what is relevant to individuals in rich countries is not the quality of typical aid programs, or even of average aid programs. What really matters is how effective the best aid programs that we can identify are. As an analogy: suppose that a group of people were championing turmeric ingestion as an antidepressant, and pointing to empirical evidence that supported their view (e.g. Ng et al 2017). How much would we learn about the effectiveness of that particular intervention by discussing how often herbal remedies work in general? Not much.

Luckily for us, GiveWell has conducted years of extensive investigation to try to find outstanding global health and development charities. Though it's impossible to tell whether the charities they've found are the 'best' charities, they have certainly found charities that do a huge amount to benefit the poor, are well-evidenced, and transparent.

Consider, for example, the Against Malaria Foundation, one of GiveWell's most highly recommended charities. A summary of the case for its positive impact is as follows:[6]

1. There is highly extensive evidence that using long-lasting insecticide treated bednets (LLIN) decreases the incidence of malaria and therefore of illness (such as anemia and splenomegaly) and death.[7]
2. These positive direct impacts come at very low costs, perhaps as low as just a few thousand dollars per life saved.
3. AMF focuses on countries that have high rates of malaria, and ensures that their distribution partners conduct post-distribution surveys (at 6-month in-

6. See GiveWell (2018a) for a much fuller analysis.

7. Summaries of this evidence can be found in two Cochrane meta-analyses, Lengeler (2004), which reviews 22 randomised controlled trials, and Gamble (2007), which reviews 6 randomised controlled trials) and on GiveWell (2018b).

tervals for a period of 2.5 years) to ensure that LLINs reach their intended destination and are being used properly.

What about possible negative impacts? GiveWell have looked into many of those, too. Are bednets used for alternative purposes, such as fishing—inadvertently depleting fish stocks due to the nets' small holes— as suggested by a popular *New York Times* op-ed? (Gettleman 2015) GiveWell have addressed this issue (GiveWell 2015), and, because AMF conduct post-distribution surveys, we can be confident that over 80% of nets are hung up 6 months after distribution. The evidence from the *New York Times* op-ed, in contrast, is based on a single study of household nets distributed along Lake Tanganyika, and provides no evidence that this is connected to depletion of fish stocks.

Might funding of AMF just displace efforts to distribute bednets by other actors? This holds to some extent, and should decrease to some extent our estimate of AMF's cost-effectiveness (and this is taken into account in GiveWell's analysis). But the bednet gap across sub-Saharan Africa is very large (requiring hundreds of millions of dollars to fill), and there are a number of clear examples of distributions that AMF have funded that would not have happened otherwise.[8]

Does saving lives lead to overpopulation? GiveWell commissioned an extensive report on this question (Roodman 2014). It is probably true that saving lives has a mild effect of increasing population size, but it's very unclear whether this should be regarded as a good, bad or neutral thing. Indeed, answering this question requires answering notoriously recalcitrant problems in population ethics.

To be clear: We cannot be 100% certain that AMF is doing more good than harm; nor can we be confident about exactly how much good it's doing (a fact that GiveWell repeatedly emphasised (GiveWell 2017)). Indeed, there are many issues that remain open. LLINs probably do contribute to insecticide resistance, though the malaria community still recommends the use of LLINs (GiveWell 2013).[9] It's possible that distributing free bednets undermines local markets for nets because people expect that they will receive them for free (see GiveWell 2012 for discussion). It's possible that AMF diverts some skilled labour from other areas, so the full costs of bednet distribution are larger than they might otherwise be (this is discussed in GiveWell 2018a).

8. See GiveWell (2018c). For discussion of how GiveWell accounts for funging with other actors in general, see GiveWell (2018d).

9. In addition, two recent randomized controlled trials suggest that next generation nets can successfully mitigate resistance: see Tiono et al (2018) and Protopopoff et al (2018).

We would be naive indeed if we did not appreciate the fact that development efforts operate in an incredibly complex context of multiple interrelated activities.

But we cannot be certain, for any of our actions, that we will do more good than harm. We should be careful to hold development efforts to the same standard we hold other activities: neither blithely assuming that any good intention will result in good actions, nor being so timorous that any possibility of harm results in paralysis. The best we can do is assess the evidence as best we can. In the case of AMF, there is a very strong case that the direct effects are very positive, a well-established track record of similar interventions being very effective, and no solid extant case for there being significant negative long-run effects.

In sum, Temkin provides no evidence at all for the idea that charities like AMF are doing more harm than good. In contrast, GiveWell provides extremely extensive evidence in support of the idea that AMF are doing much more good than harm. Given that Peter Singer, Temkin's primary target, endorses GiveWell, it's surprising that Temkin takes no time at all to engage with the hundreds of thousands of hours of work that GiveWell have invested to make their recommendation, especially when they often deal directly with the worries that Temkin has.

SECTION III

Temkin does anticipate something similar to this response. He asks: "Why can't Deaton simply support Effective Altruism? Instead of claiming that we shouldn't be supporting international aid groups operating directly in the world's poorest regions, why shouldn't Deaton contend, more modestly, that we must be very careful about which aid groups we support, to make sure that they are, indeed, doing more good than harm?" That is: he raises the idea that what is relevant ethically, is not how effective *typical* aid charities are, but instead how effective the best charities we can find are.

In response, Temkin claims that this leads us to an each-we dilemma. He claims that, by supporting those charities that effective altruists promote (such as AMF), we may each individually do the morally best thing, but we still collectively do more harm than good.

But this is clearly a non sequitur. Temkin's each-we dilemma can only get off the ground if he supposes that (i) the total aggregate harm of a single action that aims to help the poor is greater than the total aggregate benefit of that action; and that (ii)

morally, the aggregate of many small harms can never outweigh a small number of large harms (at least, for some magnitude of benefit and harm).

In other work (MacAskill forthcoming) I have suggested that effective altruism, as a matter of definition, should be committed to anonymity (roughly: that the identities of individuals across outcomes do not matter morally), and in this volume Clark and Pummer argue—correctly in my view—that anonymity is incompatible with (ii).

What I have argued in this article is that, for the charities that GiveWell recommend, we also have active reason to think that (i) is false. And if (i) is false then the total aggregate benefit of a single action is greater the total aggregate harm of that action, and so the total benefit of a million such actions will still be greater than the total harm of a million such actions.

Now, one can *conceive* of models on which there are each-we dilemmas even though any individual action does more good than harm. Perhaps, when considering a large number of actions, the goods are additive but the harms are multiplicative. But I see no reason why this should be so, I know of no empirical evidence suggesting that this is so, and Temkin gives us no reason to think this either.

CONCLUSION

Let me end with a comment about the nature of the broader dialectic regarding Singer's argument for the conclusion that we in rich countries have strong duties of beneficence. Often, critics of Peter Singer focus on whether or not aid is effective. But that is fundamentally failing to engage with core of Singer's argument. Correctly understood, that argument is about the ethics of buying luxury goods, not the ethics of global development. Even if it turned out that every single development program that we know of does more harm than good, that fact would *not* mean that we can buy a larger house, safe in the knowledge that we have no pressing moral obligations of beneficence upon us. There are thousands of pressing problems that call out for our attention and that we could make significant inroads on with our resources. Here is an incomplete list of what $10,000 can do (noting, in each case, that any cost-effectiveness estimates are highly uncertain, with large error bars, and refer to expected value):[10]

- Spare 20 years' worth of unnecessary incarceration, while not reducing

10. For recommendations for individual donors for all these cause areas, see Open Philanthropy (2017a).

public safety, by donating to organisations working in criminal justice reform (Open Philanthropy Project 2017b).

- Spare 1.2 million hens from the cruelty of battery cages by donating to corporate cage-free campaigns (Open Philanthropy Project 2016).
- Reduce the chance of a civilisation-ending global pandemic by funding policy research and advocacy on biosecurity issues (Open Philanthropy Project 2014).[11]
- Contribute to a more equitable international order by funding policy analysis and campaigning.[12]

In order to show that Singer's argument is not successful, one would need to show that for *none* of these problems can we make a significant difference at little moral cost to ourselves. This is a very high bar to meet. In a world of such suffering, of such multitudinous and variegated forms, often caused by the actions and policies of us in rich countries, it would be a shocking and highly suspicious conclusion if there were simply nothing that the richest 3% of the world's population could do with their resources in order to significantly make the world a better place.

The core of Singer's argument is the principle that, if it is in our power to prevent something very bad from happening, without thereby sacrificing anything morally significant, we ought, morally, to do so. We can. So we should.

REFERENCES

Barder, Owen. "Can aid work?: written testimony submitted to the House of Lords." *Washington, DC: Center for Global Development* (2011).

Burnside, Craig, and David Dollar. "Aid, policies, and growth." *American Economic Review* 90, no. 4 (2000): 847-868.

11. I don't provide a cost-effectiveness estimate here, but if we assume that $10 billion of carefully targeted philanthropy would reduce extinction risk by 0.1 percentage points over the coming century, that would mean that $10,000 gives a one in a billion chance of preventing an extinction-level event. That's an expected 7 present-day lives saved, and a prevention of the expected loss of life of five million future lives who otherwise wouldn't have been born; this latter estimate assumes constant population levels until the Earth is no longer habitable in 600 million years.

12. For example by donating to the Center for Global Development: see Open Philanthropy Project (2016b). Again I don't provide a cost-effectiveness estimate, but Open Philanthropy believe that over their lifespan CGD have caused billions of dollars worth of value for very poor countries on the basis of a $150 million budget.

Clemens, Michael A., Steven Radelet, Rikhil R. Bhavnani, and Samuel Bazzi. "Counting chickens when they hatch: Timing and the effects of aid on growth." *The Economic Journal* 122, no. 561 (2012): 590-617.

Deaton, Angus. *The great escape: health, wealth, and the origins of inequality*. Princeton University Press, 2013.

Easterly, William. 'Some cite good news on aid.' *Aid Watch* (2009), accessible at http://www.nyudri.org/aidwatcharchive/2009/02/some-cite-good-news-on-aid [Accessed 22/5/19].

Gamble, Carol, Paul J. Ekwaru, Paul Garner, and Feiko O. Ter Kuile. "Insecticide-treated nets for the prevention of malaria in pregnancy: a systematic review of randomised controlled trials." *PLoS Medicine* 4, no. 3 (2007): e107.

Gettleman, Jeffrey. "Meant to keep malaria out, mosquito nets are used to haul fish in." *New York Times* 24 (2015).

Givewell. 'Giving cash versus giving bednets.' www.givewell.org (2012), accessible at https://blog.givewell.org/2012/05/30/giving-cash-versus-giving-bednets/ [Accessed 22/5/19].

———. 'A Conversation with Abraham Mnzava on October 18, 2013.' www.givewell.org (2013), accessible at https://www.givewell.org/files/conversations/Abraham%20Mnzava10-%2018-13%20%28public%29.pdf [Accessed 22/5/19].

———. 'Putting the problem of bed nets used for fishing in perspective.' www.givewell.org (2015), accessible at https://blog.givewell.org/2015/02/05/putting-the-problem-of-bed-nets-used-for-fishing-in-perspective/ [Accessed 22/5/19].

———. 'How GiveWell uses cost-effectiveness analyses.' www.givewell.org (2017), accessible at https://blog.givewell.org/2017/06/01/how-givewell-uses-cost-effectiveness-analyses/ [Accessed 22/5/19].

———. 'Against Malaria Foundation.' www.givewell.org (2018a), accessible at https://www.givewell.org/charities/amf [Accessed 22/5/19].

———. 'Mass Distribution of Long-lasting Insecticide-Treated Nets.' www.givewell.org (2018b), accessible at https://www.givewell.org/international/technical/programs/insecticide-treated-nets [Accessed 22/5/19].

———. 'Counterfactual in countries where AMF funded distributions.' www.givewell.org (2018c), accessible at https://www.givewell.org/charities/against-malaria-foundation/supplementary-information#Counterfactual_in_countries_where_AMF_funded_distributions [Accessed 22/5/19].

———. 'Revisiting leverage.' www.givewell.org (2018d), accessible at https://blog.givewell.org/2018/02/13/revisiting-leverage/ [Accessed 22/5/19].

Glennie, Jonathan, and Andy Sumner. "The $138.5 billion question: when does foreign aid work (and when doesn't it)." *CGD Policy Paper 49* (2014).

Global IDEA Scientific Advisory Committee. "Health and economic benefits of an accelerated program of research to combat global infectious diseases." *Canadian Medical Association Journal* 171, no. 10 (2004): 1203-1208.

Lengeler, Christian. "Insecticide-treated bed nets and curtains for preventing malaria." *The Cochrane Library* (2004).

MacAskill, William. *Doing good better: Effective altruism and a radical new way to make a difference.* Guardian Faber Publishing, 2015.

———. (forthcoming). "The definition of effective altruism," in Greaves and Pummer (eds) *Effective Altruism: Philosophical Issues*, OUP.

Mekasha, Tseday Jemaneh, and Finn Tarp. "Aid and growth: What meta-analysis reveals." *The Journal of Development Studies* 49, no. 4 (2013): 564-583.

Ng, Qin Xiang, Shawn Shao Hong Koh, Hwei Wuen Chan, and Collin Yih Xian Ho. "Clinical use of curcumin in depression: A meta-analysis." *Journal of the American Medical Directors Association* 18, no. 6 (2017): 503-508.

Open Philanthropy Project, 'Biosecurity,' (2014), accessible at

https://www.openphilanthropy.org/research/cause-reports/biosecurity [Accessed 22/5/19].

———. 'Effectiveness of corporate campaigns,' (2016a), accessible at https://www.openphilanthropy.org/focus/us-policy/farm-animal-welfare/humane-league-corporate-cage-free-campaigns#Effectiveness_of_corporate_campaigns [Accessed 22/5/19].

———. 'Center for Global Development — General Support 2016,' (2016b), accessible at https://www.openphilanthropy.org/focus/global-health-and-development/miscellaneous/center-global-development-general-support-2016 [Accessed 22/5/19].

———. 'Suggestions for Individual Donors from Open Philanthropy Project Staff - 2017.' (2017a), accessible at https://www.openphilanthropy.org/blog/suggestions-individual-donors-open-philanthropy-project-staff-2017 [Accessed 22/5/19].

———. 'Criminal Justice Reform Strategy: What we're doing and why.' (2017b), accessible at https://www.openphilanthropy.org/focus/us-policy/criminal-justice-reform/criminal-justice-reform-strategy#What_were_doing_and_why [Accessed 22/5/19].

Protopopoff, Natacha, Jacklin F. Mosha, Eliud Lukole, Jacques D. Charlwood, Alexandra Wright, Charles D. Mwalimu, Alphaxard Manjurano et al. "Effectiveness of a long-lasting piperonyl butoxide-treated insecticidal net and indoor residual spray interventions, separately and together, against malaria transmitted by pyrethroid-resistant mosquitoes: a cluster, randomised controlled, two-by-two factorial design trial." *The Lancet* (2018): 1577-1588.

Rajan, Raghuram G., and Arvind Subramanian. "Aid and growth: What does the cross-country evidence really show?." *The Review of Economics and Statistics* 90, no. 4 (2008): 643-665.

Ravallion, Martin. "On the role of aid in The Great Escape." *Review of Income and Wealth* 60, no. 4 (2014): 967-984.

Roodman, David. "The impact of life-saving interventions on fertility,' https://davidroodman.com (2014), accessible at https://davidroodman.com/blog/2014/04/16/the-mortality-fertility-link/ [Accessed 22/5/19].

Tiono, Alfred B., Alphonse Ouédraogo, Daouda Ouattara, Edith C. Bougouma, Sam Coulibaly, Amidou Diarra, Brian Faragher et al. "Efficacy of Olyset Duo, a bednet containing pyriproxyfen and permethrin, versus a permethrin-only net against clinical malaria in an area with highly pyrethroid-resistant vectors in rural Burkina Faso: a cluster-randomised controlled trial." *The Lancet* (2018).

First Steps Towards an Ethics of Robots and Artificial Intelligence

JOHN TASIOULAS

King's College London

ABSTRACT

This article offers an overview of the main first-order ethical questions raised by robots and Artificial Intelligence (RAIs) under five broad rubrics: functionality, inherent significance, rights and responsibilities, side-effects, and threats. The first letter of each rubric taken together conveniently generates the acronym FIRST. Special attention is given to the rubrics of functionality and inherent significance given the centrality of the former and the tendency to neglect the latter in virtue of its somewhat nebulous and contested character. In addition to exploring some illustrative issues arising under each rubric, the article also emphasizes a number of more general themes. These include: the multiplicity of interacting levels on which ethical questions about RAIs arise, the need to recognise that RAIs potentially implicate the full gamut of human values (rather than exclusively or primarily some readily identifiable sub-set of ethical or legal principles), and the need for practically salient ethical reflection on RAIs to be informed by a realistic appreciation of their existing and foreseeable capacities.

1. INTRODUCTION

For almost all of human history, robots existed only as imaginary beings endowed with a Jekyll-and-Hyde character. In one guise, promising to usher in a utopia free of illness, poverty, and the drudgery of work; in another, intent on enslaving or destroying humankind. Only in the middle of last century, however, did robots achieve a

significant real-world presence when General Motors installed a robot, 'Unimate', in one of its plants to carry out manual tasks—such as welding and spraying—that were deemed too hazardous for human workers.[1] Today, robots are so commonplace in manufacturing that they are a major cause of unemployment in that sector.[2] But the use of robots in factories is only the beginning of a 'robot revolution'—itself part of wider developments powered by the science of Artificial Intelligence (AI)—that has had, or promises to have, transformative effects on all aspects of our lives.

Robots are now being used, or being developed for use, in a vast array of settings. Driverless cars have already been invented and are expected to appear on our roads within a decade. These cars have the potential to reduce traffic accidents, which currently claim more than a million lives each year worldwide, by up to 90%, while also reducing pollution and traffic congestion (Bonnefon, Shariff, Rhawan 2006). Robots are also used to perform domestic chores, including vacuuming, ironing, and walking pets. In medicine and social care, robots surpass doctors in diagnosing certain forms of cancer or performing surgery, and they are used in therapy for children with autism or in the care of the elderly. Tutor robots already exist, as do social robots that provide companionship, or even sex. In the business world, AI figures heavily in the stock market, where computers make most decisions automatically, and in the insurance and mortgage industries. Even the recruitment of human workers is turning into a largely automated process, with many rejected job applications never being scrutinized by human eyes. AI-based technology, some of it robotic, also plays a role in the criminal justice system, assisting in policing and decisions on bail, sentencing, and parole. The development of autonomous weapons systems (AWSs), which select and attack military targets without human intervention, promises a new era in military defence. And this is just a sample of recent developments.

In this article, I examine some of the key ethical questions posed by robots and AI (or RAIs, as I shall refer to them). The overall challenge, of course, is to harness the benefits of RAIs while responding adequately to the risks we incur in doing so. The need to balance benefit and risk is a recurrent one in the history of technological advance, but RAIs present it in a new and potentially sweeping form with large-scale implications for how we live among others—in relation to work, care, education, play, friendship, love—and even regarding how we understand what it is to be a

1. http://my.ilstu.edu/~kldevin/Introduction_to_robotics2/Introduction_to_robotics6.html

2. By 2012, for example, there were approximately 1,563 robots per worker in Japan's automotive industry (the figure for Germany was around 1,133 robots per worker), see Furman and Seamans (2018, 8).

human being and whether we should deploy these new technologies in the pursuit of 'human enhancement' or even, as with 'trans-humanism', in order to transcend our human condition. Prior to addressing these matters, we must first clarify some key notions. [3]

2. WHAT IS A ROBOT? WHAT IS ARTIFICIAL INTELLIGENCE?

A recent UNESCO report describes robots as artificial beings with four characteristics:

- *mobility*, which is important for functioning in human environments like hospitals and offices;
- *interactivity*, made possible by sensors and actuators, which gather relevant information from the environment and enable a robot to act upon this environment;
- *communication*, made possible by computer interfaces or voice recognition and speech synthesis systems; and
- *autonomy*, in the sense of an ability to 'think' for themselves and make their own decisions to act upon the environment, without direct external control (UNESCO 2017: 4).

The sophisticated robots that are our topic in this article operate on the basis of 'Artificial Intelligence' (AI). In the words of AI pioneer Marvin Minsky, this is 'the science of making machines do things that would require intelligence if done by men', such as face recognition or language translation. In understanding AI, two distinctions are important: (a) general and narrow AI, and (b) top-down and bottom-up AI. The first distinction relates to the scope of AI capabilities, the other to AI's technical functioning.

General AI refers to intelligent machines that are able to replicate a broad range of human intellectual capacities, and even to surpass them. These forms of AI, although familiar from science fiction characters such as *Star Wars*' C3PO, lie at best in the very remote future. To the extent that there has been any significant progress in AI in recent years, it has occurred in *narrow AI*. These are machines that replicate, or exceed, human capabilities with respect to a limited range of tasks, e.g. car-driving, medical diagnosis or language translation.

3. For helpful overviews of developments robotics and AI, and the ethical issues they raise, Edmonds (2017) and Tegmark (2017, esp ch.3).

AI operates by means of algorithms, which are rules or instructions for the solution of various problems that are usually embedded in a computer, and for present purposes can be roughly grouped into two broad kinds, corresponding to two kinds of robots. *Top-down* (or deterministic or closed-rule) algorithms control a robot's behavior by means of a pre-determined program, with the result that the robot's behavior is highly predictable. Such algorithms have been used in the preparation of income tax forms and in certain kinds of automated medical diagnoses. *Bottom-up* (or stochastic) algorithms, by contrast, enable a robot to 'learn' from past experience and revise its algorithm over time (UNESCO 2017, 4, 17-19). An illustration of this 'machine learning' is Google's DeepMind algorithm, which taught itself to play Atari games such as Breakout, inventing new score-maximizing strategies that took its own programmers by surprise. Other examples are to be found in driverless cars, facial recognition systems used by police, and algorithms recommending items to buy based on one's purchasing history. There are different kinds of 'machine learning'. Some deploy 'neural networks', which are processing nodes connected to one another in layers and modelled on the functioning of the human brain. Robots of this second sort enjoy a level of 'autonomy' not only in the sense that their behavior need not depend on human decision-making, or may not be subject to human intervention or control, but in the more radical sense that it is not readily predictable by human beings.

We should, of course, exercise caution when throwing around terms like 'intelligence', 'reasoning', 'decision' and 'autonomy' in relation to AI. These terms must not obscure the fact that a vast chasm still separates RAIs and human beings. AI systems process information as a means of recognizing patterns and relations among symbols that enable certain problems to be solved. But they cannot (as yet) *understand* in any meaningful sense what these symbols stand for in the real world (Tegmark 2017, ch.3). Moreover, even if RAIs can be successful in achieving complex goals—like recognizing a face in a crowd or translating a document from one natural language into another—they lack anything like the human capacity to deliberate about what their ultimate goals ought to be. For some philosophers, this power of rational autonomy is the source of the special dignity that inheres in human beings and differentiates them from non-human animals. No RAIs known to us, or that are realistically foreseeable, are anywhere near exhibiting such rational autonomy. [4]

4. For some scepticism about the hype surrounding Artificial Intelligence, by one of the world's leading computer scientists, see Jordan (2018).

3. ETHICAL QUESTIONS: FRAMES AND LEVELS

RAIs generate a variety of ethical questions which arise on at least three interconnected levels. One level concerns the *laws* that should be enacted to govern RAI-related activities. These laws are public standards that purport to be morally binding on all citizens in virtue of their formal enactment and which are standardly backed up by institutional enforcement mechanisms including, at the limit, punishments such as fines and imprisonment. One set of questions here concerns whether some particular aspect of RAIs should be subject to legal regulation at all; another set of questions concerns the extent to which we need to fashion specific laws to address issues thrown up by RAIs, as opposed to relying on more general legal standards. Do we need special traffic laws for driverless cars? How should the law on insurance and accident liability apply to them? Should there be criminal laws prohibiting certain kinds of robots or AI applications? In addition to domestic legislation on such matters, RAIs also raise pressing questions that require regional or international legal solutions, e.g. through treaties to outlaw AWSs or prevent the outbreak of an arms race in relation to them.

At a second level are questions about the kind of *social morality* that we should strive cultivate in relation to RAIs. This is a recognition of the fact that not all of the socially entrenched standards that properly govern our lives are, or should be, legal standards. We rely not only on the law to discourage people from wrongful behavior, such as murder or theft, but also on moral standards that are instilled in us from childhood and reinforced by society through informal mechanisms such as criticism and other extra-legal sanctions. Arguably, the very efficacy of legal regulation would be severely diminished if it could not rely upon a sustaining underlying ethical culture. Accordingly, we need to reflect on the shape of a morally sound culture in relation to RAIs.

At a third level, there are questions that arise for *individuals and associations* (e.g. businesses, universities, professional bodies, etc.) regarding their engagement with RAIs. Whatever social modes of regulation exist on these matters, individuals and associations will still need to exercise their own moral judgement. This may be because existing law and social morality lag behind technical developments, or because they are deficient in some way, or because they confer on individuals the leeway to make their own decisions on some matters. For those corporations which are at the cutting-edge of developments, the fast-changing and transformative char-

acter of RAIs may justify the elaboration of their own codes of ethics on these topics. Meanwhile, others have called for a 'Hippocratic oath' for data scientists to establish an ethical framework for their operations independently of applicable legal standards (Upchurch 2018).

Difficult questions arise as to how best to integrate these three modes of regulating RAIs, and there is a serious worry about the tendency of industry-based codes of ethics to upstage democratically enacted law in this domain, especially given the considerable political clout wielded by the small number of technology companies that are driving RAI-related developments. However, this very clout creates the ever-present danger that powerful corporations may be able to shape any resulting laws in ways favourable to their interests rather than the common good (Nemitz 2018, 7). Part of the difficulty here stems from the fact that three levels of ethical regulation interrelate in complex ways. For example, it may be that there are strong moral reasons against adults creating or using a robot as a sexual partner (third level). But, out of respect for their individual autonomy, they should be legally free to do so (first level). However, there may also be good reasons to cultivate a social morality that generally frowns upon such activities (second level), so that the sale and public display of sex robots is legally constrained in various ways (through zoning laws, taxation, age and advertising restrictions, etc.) akin to the legal restrictions on cigarettes or gambling (first level, again). Given this complexity, there is no a priori assurance of a single best way of integrating the three levels of regulation, although there will nonetheless be an imperative to converge on some universal standards at the first and second levels where the matter being addressed demands a uniform solution across different national jurisdictional boundaries.

Deepening this complexity is the fact that the fields of AI and robotics are both rapidly changing and the focus of considerable hype, making it hard to disentangle realistic future scenarios from mere science fantasy. In light of this, our ethical thinking at all three levels must be sensitive to the time-frame in question, sometimes addressing matters of immediate concern, other times anticipating future developments. A persistent danger is that we are distracted by potential developments that will arise, at best, in the very remote future, while neglecting pressing concerns in the here and now. In what follows, an attempt will be made to keep the focus on the here and now, as well as realistic future scenarios, although inevitably more speculative scenarios will also be broached.

4. FIVE MAJOR MORAL ISSUES—A F*I*R*S*T ANALYSIS

Many, if not all, of the moral questions raised by RAIs can be arranged under five main headings—functionality, inherent significance, rights and responsibilities, side-effects, and threats—with the first letter of each rubric conveniently generating the acronym 'FIRST'. Of course, the boundaries between the five distinct headings are not always sharp, and although I will usually refer to RAIs compendiously as a group, different kinds of RAIs will raise significantly different kinds of concerns under each of these five headings. The acronym is apposite because the issues discussed below are first-order questions about the rights and wrongs of our engagement with RAIs. There are, in addition, important second-order questions, regarding the procedures we should adopt in addressing these first-order issues, such as standards of transparency or democratic accountability. But these second-order matters are largely beyond the scope of this article. In what follows I focus primarily on functionality and inherent significance, giving only highly compressed treatments of the other three headings.

4.1 FUNCTIONALITY

The first issue is whether a proposed RAI, e.g. a driverless car, is functional. I take 'functionality' here in an expansive sense that is not neutral with respect to the moral quality of the ends pursued by a RAI or the means it adopts in pursuing them. Functionality concerns a RAI's ability to: (a) achieve a *worthwhile* goal, e.g. transporting passengers to their desired destination, and to do so: (b) *effectively* i.e. with a reliable degree of success, (c) *efficiently* i.e. without undue expenditure of resources, and (d) in a *morally appropriate way* i.e. without violating moral norms as an inherent part of its operation, irrespective of the intent of the designer, e.g. rights to life or privacy or norms of environmental protection. Although all these dimensions raise important questions, let us focus on the last one, which throws up two large questions: (1) what are the moral standards that apply to RAIs?, and (2) how can they be built into the operation of RAIs?

A famous attempt to address the first question is Isaac Asimov's 'Three Laws of Robotics':

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.

2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Asimov 1950, 40).

But Asimov's laws immediately run into problems. One is a lack of clarity about the concepts of 'injury' and 'harm' in the first law. If a robot bodyguard injures a would-be assassin in the course of protecting an innocent person, has it 'injured' or 'harmed' him? The interests of the assassin have obviously been impaired, but has he been wronged? We need to distinguish between non-moralized and moralized conceptions of harm or injury (or, in the legal version of the distinction, *damnum* and *injuria*). When we do so, it seems unlikely that a complete ban on RAIs harming human beings in the non-moralized sense will be sustainable. Indeed, even requiring that RAIs never wrong a human being may underestimate the complexity of the dilemmas RAIs may legitimately confront.[5] A familiar dilemma concerns how a self-driving car should respond to situations where there is a choice between avoiding harm to its passenger—e.g. by swerving out of the path of an oncoming truck—versus avoiding harm to other humans (drivers, passengers or pedestrians) who are at risk of death or injury if the car swerves to save its passenger. This 'trolley problem' receives conflicting responses, but any plausible answer seems to entail the all-things-permissibility of a wrong being done to a human being by a RAI. Interestingly, empirical studies indicate that most people agree that passengers should be sacrificed in order to save a greater number of bystanders, yet most would also prefer to ride in a car that always save its passenger (Bonnefon, Shariff, Rahwan 2006).[6] If so, identifying the correct answer to the trolley problem may turn out to be practically irrelevant, as not enough people would buy the first type of car to make it worth producing. Again, Asimov's second and third laws may be questioned if we are persuaded that advanced RAIs, with seemingly human-like intellectual and emotional traits, acquire something approximating human personhood and the rights, including to self-defence, that flow from it.

Asimov's principles are an early and rudimentary attempt at constructing an ethics for RAIs. But similar elementary difficulties plague more recent efforts, such

5. Cf. also the fifth framework ethical principle outlined in House of Lords AI Committee 2018, 125: 'The autonomous power to hurt, destroy or deceive human being should never be vested in artificial intelligence'.

6. For a discussion of ethical issues related to how self-driving cars should handle accidents, see Nyholm 2018.

as the Asilomar AI Principles formulated in 2017. Some of these principles verge on the truistic, e.g. principle 6 which requires that 'AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible'. Others are unexceptionable at the price of being unhelpfully vague, e.g. principle 15 which states that 'the economic prosperity created by AI should be shared broadly, to benefit all of humanity'. Two other tendencies of the principles are worth highlighting, since they are often replicated in other declarations of ethical principles for RAIs. The first is the implicit assumption that there is an enumerable catalogue of evaluative considerations that are especially engaged by RAIs. Thus, principle 11 demands compatibility of AI systems with 'ideals of human dignity, rights, freedoms, and cultural diversity'. But it is questionable that any meaningfully specific list of RAI-salient values is in order. Why not include additional values such as charity, respect for the natural environment or concern for the common good, among others? There is no reason *ab initio* to suppose that the ethical values potentially applicable to RAIs fall short of the entire range of human values. There is, of course, some recognition of this when the principles invoke other values, such as the common good. But here a second worrying feature crops up, which is the tendency to reduce the enumerated values to widely held *beliefs* about value. Hence, principle 23 on the common good states: 'Superintelligence should only be developed in the service of widely shared ethical ideals, and for the benefit of all humanity rather than one state or organization'. There is a conflation here of two distinct notions of the common good: (1) ethical values that are in fact widely shared among human beings, and (2) that which would objectively benefit all human beings. The latter is a normative idea, the former an empirical one whose normative implications, if any, need to be worked out in tandem with genuinely normative principles. The problem does not go away if an appeal is made to law, rather than widely held beliefs. The European Commission's recently published *Ethics Guidelines for Trustworthy AI*, for instance, accord a foundational role to human rights law.[7] Leave aside the fact that this law does not reflect all of the ethical considerations (e.g. environmental values) bearing on AI, that it does not tend to be directly binding on non-state actors, and that not all of its provisions bind all states (e.g. because they have not ratified relevant human rights treaties). The

7. "We believe in an approach to AI ethics based on the fundamental rights enshrined in the EU Treaties, the EU Charter and international human rights law. Respect for fundamental rights, within a framework of democracy and the rule of law, provides the most promising foundations for identifying abstract ethical principles and values, which can be operationalised in the context of AI". European Commission 2019, 9.

more fundamental point is that such laws—despite the powerful moral charge conferred by the words 'human rights'—are not basic ethical standards. Instead, like any other set of laws, they are themselves to be formulated and evaluated—and, sometimes, to be found seriously wanting—in terms of basic ethical standards, including the morality of human rights (see Tasioulas (forthcoming)).

A sound ethical approach to RAIs, therefore, must go beyond invoking widespread beliefs or established law, including human rights law, to engage the full gamut of relevant ethical values. The tendency to conflate the normative with the empirical, conventional or legal is perhaps to be expected among those with a technological and 'data-driven' mind-set, who for understandable reasons tend to be prevalent in the robotics and AI community. It can lead them to the disastrous conclusion that ethical standards are to be identified through the deployment of empirical methods—for example, 'crowd-sourcing'—for ascertaining widespread ethical beliefs. Down this path, however, lies the degeneration of ethics into a branch of the public relations industry.

The second question, about how to build ethical norms into the workings of RAIs, is no less challenging. Some entertain high ambitions for robot moral sages commanding an expert knowledge of morality that far surpasses the average human. Julian Savulescu and Hannah Maslen contend that 'artificial ethical agents', thanks to their superhuman speed, extensive data-bases, and lack of characteristic human vices, such as selfishness, could 'actually aid or even replace humans when it comes to difficult moral decision-making' (Savulescu and Maslen 2015). Along similar lines, RAIs have been proposed as a means of overcoming the biases that notoriously afflict human judges in the sentencing of criminals—such as sentencing more leniently or harshly depending on the time of day or, more worryingly, in response to the class, ethnicity or race of offenders. [8]

Others, such as the authors of a recent UNESCO report, are sceptical about the moral perfectibility of RAIs.[9] Such scepticism has two inter-related sources. First, that moral decision-making confronts a potential infinity of relevantly different situations that no algorithm or process of machine learning is sensitive enough to engage

8. See, for example, Sunstein (forthcoming), on how algorithms can help correct for the Current Offender Bias—the tendency to place an excessive emphasis on the fact of the current offence—in decisions about bail. More generally, for the claim that human cognition is more of a 'black box' compared to the levels of transparency that can be potentially achieved in relation to detecting discrimination by algorithms, see Kleinberg, Ludwig, Mullainathan and Sunstein (forthcoming).

9. 'It does not seem probable that any machine that lacks emotions like empathy... could deal with this variation of morally relevant facts and preferences'. UNESCO (2017, 44).

with adequately. And, second, that sound moral reasoning requires the cultivation of emotional responses on the part of the reasoner, such as guilt, indignation, and empathy, that are properly attuned to their objects. It is these responses that enable us to register the moral significance of certain situations, e.g. the need to act with urgency in situations warranting fear of an imminent threat. But, arguably, they are inherently beyond the capacities of beings that do not share in a human consciousness and way of life. Both lines of thought have been stressed by a diversity of traditions in moral philosophy, most prominently in recent times by neo-Aristotelian virtue ethics and feminist theory. But they also receive some support from the behavioural and brain sciences, which suggest that the capacity for emotion is not a distinct 'module' that is added to our cognitive machinery, but an integral part of the overall architecture of our brains (Pessoa 2018).

Whether RAIs can become moral experts will partly depend on what is the correct philosophical account of morality. If the correct view is something like utilitarianism, which rests on a single very general ethical principle and requires daunting calculations of future consequences, the prospects for RAI sages may seem bright. If, by contrast, the correct moral approach requires context-specific judgment drawing upon a rich palette of moral emotions attuned to a plurality of values, and gives no role to mechanically applicable general principles, then the prospects seem correspondingly bleak.

Even if we set aside futuristic speculations about RAIs as moral experts, and focus instead on their compliance with basic moral norms in the performance of various specific tasks, the truth about morality will have an important bearing on how such norms are best integrated into the workings of robots. Of course, much will depend on what sort of tasks we wish robots to perform and what kind of settings they will operate in, e.g. whether this will be with or without potential human supervision or override. However, it would be unrealistic to suppose that we have to resolve disagreements in moral philosophy before we program ethical principles into the workings of RAIs. This is because proponents of different moral philosophies can still have good reason to converge on some core of basic moral standards, even if they would offer different justifications and (in many cases, interpretations) of them.

As we saw, two kinds of approaches to instilling ethical standards into the operations of RAIs can be distinguished, although elements of both approaches could be blended in any given RAI. The first is a top-down approach that involves rendering certain principles—such as the Geneva Conventions on the conduct of war, in the

case of AWSs, or norms of conversational reasonableness and non-offensiveness for 'chatbots'—into algorithmic form. This is a daunting task, one which, if successfully carried out, would enable a RAI to make sophisticated judgments of proportionality when using lethal force or to distinguish between playful humour and offensive slurs. Another approach is more bottom up in character. It would proceed on the basis of a form of 'machine learning' in which, for example, the RAI might be exposed to a vast number of past decisions made by legal experts in the relevant field and then proceeds to extrapolate to decisions of its own in future scenarios. Taking their cue from virtue ethics, some have even argued that the right way to instil ethics into RAIs is to raise them as we do children, on the basis that the development of good character requires a decent upbringing (Rini 2017).

Although RAIs promise to assist us in meeting certain challenges, including overcoming human imperfections and limitations in carrying out important tasks, they are often defective in complying with appropriate standards for achieving an otherwise valuable goal. As we noted above, AI is powered by enormous data-sets. But the means by which data is accumulated is often morally dubious. One notorious example is the targeted advertising carried out by online platforms like Facebook and Google, from which they derive around 90% of their revenue. It might be convenient, if also unnerving, to have advertisements appear in your online newsfeeds that are tailored to your tastes and interests. But this process involves algorithms operating on data gleaned from the websites you visit, the emails you send, and mobile tracking. And there is a serious question as to whether people using these platforms are aware that they are 'data cows' being relentlessly milked for commercially valuable information, let alone whether they have meaningfully consented to it. As a result, these business models have rightly sparked concerns that they violate privacy rights or constitute forms of economic exploitation. Similar concerns extend to many other activities, including platforms such as that developed by the Axon corporation, which store more than 20 million gigabytes of public safety-related data taken from police body cameras (Goode 2018).

Giving individuals the right to control what is done with their personal data is not an all-purpose solution to the ethical problems created by the gathering and use of such data. One reason for this is that data may be needed to advance a vital social good—for example, to prevent the outbreak of a contagious disease or to anticipate a terrorist attack. In such cases, giving individuals a veto over whether their data is accessed or used in certain ways seems disproportionate to the value of the good

that is forgone. However, the pursuit of social goods in this way in the absence of individual consent may require the fulfilment of other stringent conditions, such as transparency in the aims and methods of data-users, mechanisms for holding them to account, and so on.

Another defect that those working in the field have struggled to eradicate is that of algorithmic bias, which can arise even if the means of capturing data do not violate moral norms, such as those of privacy and non-exploitation. RAIs are driven by algorithms that are trained on datasets and operate by generalizing from them to future scenarios. One problem arises from the fact that the training data may itself be defective as a basis for accurate judgments or decisions. This is unsurprising, since the data are generated by the very fallible creatures (human beings) whose shortcomings, many of which are congealed in the form of harmful and unjust social patterns of behaviour, the RAIs were developed to overcome in the first place. In particular, the data may be statistically skewed, e.g. not inclusive of minority groups or embodying prejudices and historical patterns of discrimination. Recent examples of real-life algorithmic bias include an algorithm used by an English police force that discriminated against people from poorer areas in deciding whether to keep offenders in custody, job search tools that favoured men over women for high income jobs, and facial recognition algorithms, used in a range of applications from internet image searches to police enforcement, that have much higher error rates for women and non-white people (Oswald, Grace, Unwin and Barnes 2018; Burgess 2017; O'Neil 2016; and Buolamwini and Gebru 2018). A particularly egregious illustration is the risk-assessment algorithm COMPAS, which has been used by some American courts in making sentencing decisions. Its aim is to predict the likelihood of an offender re-offending within the next two years, an aim it fulfils at a 70% accuracy rate. However, its errors exhibited a pronounced racial bias: COMPAS was twice as likely to make an erroneous finding of high risk in the case of a black offender than a white offender (a 'false positive'), meanwhile, white offenders were twice as likely to be erroneously labelled as low risk as compared with black offenders (a 'false negative') (Larson, Mattu, Kirchner and Angwin 2016). [10]

Now even if, say, a risk-assessing algorithm is effective in accurately predicting the likelihood that suspects possessing certain characteristics will offend, this does

10. For a more recent discussion that places the COMPAS algorithm controversy within the wider philosophical debate about fairness and discrimination, see Binns (2018). See also, more generally, Barocas and Selbst, (2016).

not put an end to our ethical worries. For one thing, there is something problematic about a person having a judgment made about him on the basis of how *other* people, who share various of his characteristics, behaved in the past. Does not respect for his personal autonomy require that he be assessed on the basis of the distinct individual that he is, his previous history of words and deeds, rather than those of others who share his (perhaps unchosen) characteristics? The problem is aggravated in cases where the relevant characteristics, such as race, gender or poverty, are themselves the focus, or the products, of a long history of grave injustice. The result is a 'vicious cycle' in which the RAI perpetuates and exacerbates the history of injustice that generated its dataset, e.g. by focusing on minority or poor suspects in making arrests, leading to greater numbers of minority convictions, which in turns leads to greater discrimination against minority suspects, in an endless downward spiral.

This problem of algorithmic bias cannot be solved by simply winnowing out data that is explicitly about sensitive categories such as age, class, gender or race, since other seemingly innocuous types of data may correlate with (or be proxies for) these categories. One important part of the answer to such challenges is to scrutinize carefully the purposes for which an algorithm is deployed (Fry 2018, 62). For example, it may be that in some contexts what really matters is the overriding issue of predicting outcomes, whereas in other contexts this is of lesser importance. Arguably, for example, determining which suspects should be given bail allows greater scope for a decision based on the predicted likelihood of a suspect committing an offence than does a sentencing decision, since the former does not carry anything like the same level of moral opprobrium as the latter. The challenges already mentioned to the just operation of algorithms are compounded by the fact often, for commercial reasons, neither the algorithm nor the data on which it is trained are made public. Moreover, in the case of bottom-up algorithm, it can be opaque even to the people operating the RAI precisely what algorithm is governing its activity.

If a RAI is functional in these ways discussed above, a question arises as to the level at which the requirement of functionality should be pitched, especially in comparison to what human beings can achieve. The answer will depend on the value of the goal being served and the risks we should be prepared to run in order to achieve it. In the case of some tasks, such as those performed by house-cleaning or companionship robots, it may be enough that a RAI performs adequately even if not quite as well as humans. Regarding other tasks, such as driving, medical diagnosis or the sentencing of criminals, what is at stake may be judged so important—life, liberty and justice—

that RAI functionality must be at least as good, or perhaps significantly better, than that ordinarily achieved by human beings. Superior performance by RAIs is something that may also be needed to offset the unwelcome side-effects and threats that reliance on them may also generate, such as job losses or sabotage by malicious agents (see 4.4 and 4.5, below). A complicating factor in identifying the minimum acceptable level of functionality is the baseline of comparison: is it what humans are in principle capable of achieving or what they are actually likely to achieve? For example, in a world in which millions of people lack access to basic education and health care, the lessons of a robot tutor or diagnoses by a robot physician may have significant value even if they are decidedly inferior to the services of human tutors and physicians that individuals cannot access, whether for financial or other reasons. It would be perverse to deprive people of vital benefits provided by RAIs in the name of an idealized level of human service that is likely to remain beyond their reach.

A further consideration we need to keep in mind is the dynamic quality of the moral standards bearing on RAIs; in particular, the way in which they may evolve over time as a result of technological developments and the new profile of opportunities and costs that they may come to generate. What I have in mind here is not a simple-minded one-way determination of our moral standards by technological developments, the sort of view that erroneously leads some to claim that 'privacy is dead' just because there is no foolproof way of preventing violations of privacy. Rather, the idea is that moral standards which were worked out in an era prior to the emergence of capabilities associated with RAIs may need to be reassessed in light of their emergence and the benefits they promise or the risks they pose. If, for example, the development of RAIs significantly increases the possibility of predicting the outbreak of epidemics by means of digital disease detection, we may need to develop less constraining norms regarding the kinds of digital surveillance RAIs may undertake to protect us from such threats. Forms of surveillance that, in the past, may have legitimately been judged violations of the right to privacy may, in this new technological environment, no longer be so (Vayena and Tasioulas 2016).

4.2 INHERENT SIGNIFICANCE

Even if RAIs can achieve sufficient functionality in a given task or role, questions may arise about the inherent significance of assigning that task or role to RAIs and not humans. Sometimes, the elimination of the 'human factor' may be beneficial.

In the case of an elderly person who needs assistance when bathing, for example, a robot carer can minimize the risk of embarrassment.[11] But the elimination of the human factor can also be troubling. This is already evident in a widespread concern that people should know whether the 'other' they are interacting with online or over the telephone is a human being or a machine. One context where the relevance of the human factor seems especially acute is that of RAIs making decisions that have serious consequences for human beings. What is mainly at issue here are especially (1) decisions based on bottom-up or stochastic algorithms which are therefore not reliably predictable, and (2) decisions that turn upon some kind of personal evaluation of the human being who is affected, e.g. their merit, desert or entitlements.

In reflecting on this concern, it will be useful to have in front of us a vivid, if compressed and contestable, reminder of the significance of being a human, especially as contrasted with being an artefact such as a machine. In this regard, we can do little better than the following statement by David Wiggins:

> *[O]ur sharing in a given specific animal nature and a law-sustained mode of activity is integral to the close attunement of person to person in language and integral to the human sensibilities that make interpretation possible. Secondly, that sharing in a specific animal nature and mode of activity is a precondition of the human solidarity (where present) that excoriates the treatment of a human being—of one of us—as a mere thing or a mere tool. And, thirdly, that affinity in nature and activity is integral to our picture—a non-deterministic picture—of our capacity, singly and collectively, to determine, within a framework not of our own choosing and replete with meanings that are larger than we are, our direct and indirect ends. Within this framework we can find our place and exercise our capacities. We see ourselves not as things with a function—what on earth could a person, as a person, be for?—but as autonomous, self-moving, animate beings, beings who find themselves in the world and seek to leave their own mark upon it, make the best of what they find there, and look (if they are lucky) for something that each one of us can come to think of as his or her own proper work or calling (Wiggins 2016, 91).*

In this passage, Wiggins accords a threefold significance to a shared human nature: *hermeneutically*, it enables a certain kind of mutual understanding of one

11. For a thoughtful discussion of the use of RAIs in health and social care, in light how technological advances impact on human relations in the context of modernity, see Coeckelbergh (2015).

person by another, whether or not mediated by language; *ethically*, it is a basis for a certain kind of solidarity premised on a shared inherent value that is inconsistent, among other things, with the treatment of fellow humans as mere tools for the furtherance of our individual ends; and, *metaphysically*, humans have the rational autonomy to determine their life-shaping ends, as opposed to artifacts whose nature and activity is given by some fixed and specific purpose set by others.

Let us now return to our concern with the inherent significance of assigning decision-making functions to RAIs. Consider, for example, the plight of long-term unemployed people whose job applications are routinely rejected by the automated systems that now dominate workforce recruitment. After months or even years of applying unsuccessfully for jobs, those individuals may never once have their application read and evaluated by a fellow human. Even if we assume that the relevant algorithm meets a good standard of functionality, i.e. it is just as effective, efficient and compliant with norms of appropriateness as the average human recruiter, the fact that it is a non-human mode of decision-making is worrisome. It is hard to pin down the worry very precisely, but the thought is roughly that the job seeker is subjected to a cold, alienating, and ultimately potentially disrespectful process because his application never comes to the attention of a fellow human being. So much is suggested in this extract from a recent *Guardian* article:

> *"It's a bit dehumanising, never being able to get through to an employer," says Robert, a plumber in his forties who uses job boards and recruiters to find temporary work. Harry, 24, has been searching for a job for four months. In retail, where he is looking, "just about every job" has some sort of test or game, anything from personality to maths, to screen out applicants. He completes four or five tests a week as jobs are posted. The rejections are often instant, although some service providers offer time-delay rejection emails, presumably to maintain the illusion that a person had spent time judging an application that had already failed an automated screen (Buryani 2018).*

Or consider two other illustrations that engage even more vital interests: the sentencing of criminals by RAIs and the use of autonomous weapons systems (AWSs). Of course, there is a serious question as to whether RAIs can achieve a level of functionality comparable, or superior, to that of human judges and soldiers. [12] But even if

12. For a comparatively upbeat view of the possibilities here, see Turner (2018).

they can, we might be troubled by the fact that a RAI's decisions impact so drastically on human life and liberty. Now, a sceptic might ask: what justified complaint can an accused or enemy soldier possibly have about the *mere fact* that they were sentenced, or killed, by a non-human being? Is the worry here nothing more than nostalgia masquerading as an ethical qualm? Perhaps one way of articulating the concern is by means of the thought that decisions about the life and liberty of others are so significant, something of value is lost if they are not made by an agent who can take responsibility for them. And, paradigmatically, for us, that agent is a human being—someone who can understand and empathise with our plight as a fellow human and reach a decision in light of their own autonomous assessment of the reasons in play for treating us one way rather than another. Here, all three of Wiggins' dimensions of the human being are in play. A corollary of this is that there is no decision-maker of whom one can demand *their* reasons for reaching the decision they did and whom one can hold accountable given an evaluation of those reasons. There is a valuable human solidarity and reciprocity—human beings recognizing each other as fellow human beings and forming their attitudes and decisions about each other on that basis—that is lost in the context of dehumanized, fully automated decision-making.

To this we can add a further observation. It will often be the case that the relevant reasons bearing on how another should be treated can be balanced in a variety of ways, with a range of decisions being rationally eligible, but no one decision being the single correct answer. In sentencing criminals, for example, the range of sentences open to the judge will typically vary in severity within certain rough bounds, with no exact amount of punishment being determinately singled out as the one which is uniquely justified. In response to this situation, some will prize consistency in the handling of similar cases above all else, claiming that criminals who commit the same wrong must receive exactly the same amount of punishment. RAIs, they might insist, are especially well-placed to treat 'like cases alike' in this way. But others see a place for discretion on the part of the judge—and the consequent variation in sentences that this generates. For those who adopt the second approach, there is value in a merciful judge being able to express their values and their character by, for example, choosing a more lenient sentence from a range of eligible options. There is value in a criminal justice system that offers offenders the possibility of discretionary mercy. The granting of mercy here is a kind of gift-giving, by one person to another, perhaps reflecting the hope of the former that the latter has genuinely repented of their past wrong-doing. In the case of the automated sentencing system, this value would be

severely curtailed or eliminated. The robot is not an individual, with values and a character of its own, who can respond to the offender's plea for mercy as one human being to another, choosing a more lenient sentence when a harsher one is also rationally open.

Now, even if there is a generalized concern about RAIs making important decisions bearing on the interests of human beings, there are still two further issues that need to be addressed. The first is the weight that the human factor possesses in any given case. Is it simply one reason among others, or does it rise to the level of an obligation, perhaps even one associated with a human right? A report by the Rathenau Institut has proposed a human right to meaningful human contact (as well as a right not to be measured, analysed, or coached) (UNESCO 2017, 39), one that is presumably threatened in cases of the three kinds discussed above. Meanwhile, the UNESCO report previously mentioned rejects AWSs on the basis of a 'guiding principle that machines should not be making life or death decisions about humans', since delegating the decision to kill a human to a machine is an affront to human dignity (UNESCO 2017, 54).[13] Of course, even if a weighty value is being sacrificed, it may be that, all-things-considered, the benefits of robot judges or AWSs justify that sacrifice in some of these cases. More radically still, some contend that we should welcome the eventual erasure of the 'human-machine divide' through the emergence of human-machine hybrids or cyborgs (see 4.3, below). More realistically, however, even if we accord inherent significance to the human factor, its weight is liable to vary from one domain to another: perhaps in cancer diagnosis it is of no or negligible value, with everything being subordinated to the question of which is the most accurate method of diagnosis, whereas in criminal sentencing it seems intelligible to accord significance to the human factor, even if this results in an overall reduction in the soundness of sentencing decisions.

The second issue relates to ways in which concerns about the loss of the human factor might be tempered while still giving RAIs an important role. Minimally, it might be thought, those affected by the decisions of robot judges should be able to seek disclosure of the basis on which those decisions were made. [14] This last require-

13. For a useful discussion of robots designed to kill, see R. Sparrow, 'Killer Robots', *Journal of Applied Philosophy* 24 (2007); 62-77.

14. This is in line with the 'right to an explanation' some claim arises from the 'right not to be subjected to automated decision-making' in the EU's General Data Protection Regulation (2016); but for some scepticism about whether any such (practicable) right has been established, see Wachter, Mittelstadt, and Floridi (2017).

ment is more demanding than it may appear, and not just because commercial incentives lead to algorithms being shrouded in secrecy but also because creators of RAIs often candidly admit that they do not fully understand how their creations, operating on the basis of bottom-up algorithms, manage to perform in the successful ways that they do. In addition, there may be scope for human supervision and override regarding sentences passed by robot judges. Or, in the job application case, it might be that some random sample of applications is evaluated by human recruiters, so that the system is not entirely devoid of direct human input. All these suggestions head towards the general conclusion that RAIs might play a valuable role in *assisting* humans to make important decisions rather than completely replacing them in performing such tasks. But it is doubtful that this could serve as an all-purpose solution, since there might be scenarios, such as battlefield decisions by AWSs, where comprehensive human supervision is simply incompatible with securing the benefits promised by RAIs.

I have so far focused on one broad category of cases in which the inherent fact that a RAI is performing a function has potential ethical significance: that in which an authoritative decision is made that impacts severely on human interests. But concern about the human factor also arises in other sorts of cases, not least those in which RAIs are treated as partners in relationships of friendship or intimacy. The absence of the human factor here can provoke a feeling of creepiness and revulsion. Its source lies in the introduction of a robot, which is ultimately a machine or instrument, into a form of relationship that—up until now—we have reserved for humans. Some will respond that the 'human factor' has here degenerated into a visceral 'yuck factor' and carries no more normative weight than the revulsion homophobes or racists feel towards same-sex or mixed-race couples. But such a dismissal would be too quick. Ideally, in a romantic relationship, the parties value each other not simply as means, but as ends in themselves, partly in virtue of their ability freely to reciprocate feelings and emotions such as love and affection. In this context, dating a robot with a human form may not be quite as absurd or weird as a romantic attachment to an iPhone or smart fridge, but it's arguably on a continuum with them.

Imagine now a social world in which large numbers of people prefer RAIs over humans as their friends and romantic partners, perhaps because they are programmable to the user's own specifications and lack the independence of mind that can be a source of friction, disappointment, and betrayal in the case of human friends and lovers. But the very word 'user' here highlights what such a relationship lacks

as compared with the value potentially available in a personal relationship between autonomous human beings. An instrument is being made to play a role that belongs to a person. That the instrument may have certain good effects on the happiness or state of mind of the user does not erase the fact that it remains a *mere* instrument and not a person. The flip side of this point is the feeling that someone who prefers robot friends to human friends has not truly succeeded in growing up. It is somewhat as if a middle-aged man still treated his childhood teddy bear as his 'best friend'.

Of course, even if you agree that something of value is lost here, nothing automatically follows about whether people should choose relationships with robots, or whether such relationships should be condemned by our social morality or prohibited by law. After all, it may be that a relationship with a robot is the only kind of relationship realistically available to some people or that the loss of the human factor is compensated by other benefits the relationship brings. Or it may be that we should simply respect people's free choice to enter into such relationships even if we judge these relationships to be deeply deficient. But the starting-point for addressing any of these questions is deeper reflection on the value of something we have largely taken for granted: the presence of the human factor in our everyday lives. [15]

4.3 RIGHTS AND RESPONSIBILITIES

If a RAI is functional and there are no compelling inherent reasons against deploying it, questions arise as to whether it possesses a moral status that confers upon it rights and responsibilities. As an artifact, created to further our ends, it seems doubtful that a RAI can possess the inherent value required to ground such a status. If RAIs came close to replicating our general capacity for rational autonomy, there would be a case for according them a comparable moral status to human beings, with corresponding rights as well as responsibilities. Indeed, it may be that in coming years the human/machine divide itself will gradually blur and even disappear with the appearance of various hybrids or cyborgs (RAIs integrated with human beings). In the words of a proponent of 'transhumanism', Ray Kurzweil:

> *Computers started out as large remote machines in air-conditioned rooms tended by white coated technicians. Subsequently they moved onto our desks, then under*

15. For explorations of some of the issues posed by the possibility of roots friends / lovers, see Danaher and McArthur Eds (2017).

our arms, and now in our pockets. Soon, we'll routinely put them inside our bodies and brains. Ultimately we will become more nonbiologial than biological (Kurzweil 2002. For a fuller discussion, Kurzweil 2005).

Such developments could have radical implications for the content of the moral standards that apply to RAIs, such as a right to self-defence that could justify killing or harming humans who pose a threat. However, as we have already seen, general AI seems a very distant prospect, even if it cannot be totally ruled out as a logical possibility.

The more pressing question is whether a good case exists for attributing legal personality to RAIs, with corresponding legal rights and responsibilities, on analogy with other 'artificial persons' recognized by law, such as corporations.[16] The issue is not that of treating RAIs the same as human beings for all legal purposes since the legal personality of RAIs need not precisely match that enjoyed by ordinary human beings, or 'natural persons'. It may consist in a different, and probably considerably smaller, bundle of rights and obligations. And the relevant bundle may vary in content from one kind of RAI (self-driving cars) to another (health care RAIs).

The case for attributing legal personality to RAIs seems more plausible in relation to RAIs operating with bottom-up algorithms, given that their behavior is not fully predictable. In the case of RAIs operating on the basis of top-down algorithms, which render their behaviour highly predictable, the argument for attributing legal responsibility to manufacturers, owner or users seems compelling. Along these lines, the European Parliament has entertained the possibility of 'creating a special legal status for robots in the long run, so that at least the more sophisticated autonomous robots could be established as having the status of electronic persons responsible for making good any damage they may cause, and possibly applying electronic personality to cases where robots make autonomous decisions or otherwise interact with third parties independently' (European Parliament 2017, para 59(f)). This proposal is reminiscent of the old English common law notion of a 'deodand', according to which animals or things, including (then) new technologies such as trains, could be forfeited if they caused damage to another person, without any fault being attributed to their owner. As a result of the forfeit, the value of the thing in question could be used to compensate anyone who had been harmed by it. The doctrine was eventually

16. For some recent discussions, Schwitzgebel and Garza (2015); Gunkel (2018); Bryson, Diamantis, and Grant, (2017); and Turner (2018).

abolished in the nineteenth century and replaced by no-fault liability and insurance schemes to compensate those harmed by newly emerging technologies.

Given the assumption that the attribution of legal personality to RAIs would not be grounded in their independent moral personhood, the question of whether RAIs should be designated legal persons falls to be determined by reference to the overall balance of benefits and burdens for human beings, and other morally considerable beings, of adopting this legal innovation (Bryson, Diamantis, and Grant, 2017). Critics have drawn attention to various pitfalls of granting RAIs legal personality. One pitfall is the possibility of abuse (a consideration belonging to the category of threats, discussed at 4.5, below): the legal personality of RAIs might be used by other unscrupulous agents, e.g. their manufacturers and operators, in order to evade any responsibilities on their part. Another arises where there is no relevant human agent 'standing behind' the actions of a RAI: how will the RAI compensate others for its breach of obligation? (Bryson, Diamantis, and Grant, 2017). Would the successful litigant be given ownership of the RAI, or might RAIs possess assets, e.g. bank accounts, that could be claimed as compensation? But these are not conclusive objections, and they certainly do not definitively rule out the advisability of limited forms of legal personality for some RAIs.

Even if there are serious obstacles in the way of according legal personality to RAIs, it is arguable that bottom-up RAIs are sufficiently different from most machines that existing law must be revised to take into account their capacity for autonomous behaviour. Along these lines, the UNESCO report proposes a threefold scheme: (1) use of existing law in assigning responsibility in relation to top-down robots; (2) for bottom-up robots, use codes of practice and ethical guidelines in addition to the law; and (3) for bottom-up robots that can harm human beings, e.g. AWSs or self-driving cars, consideration of the degree to which autonomous decisions should be left to the robot and where meaningful human control should be required (UNESCO 2017, 48-9). It is doubtful, however, that any such neat schema is ultimately defensible. The injunction in (3) seems plausible, if rather vague, but (1) is misguided. There is no reason to suppose that only legal standards are needed in the case of top-down RAIs, any more than only legal responsibilities arise for the users of a standard car.

Underlying all of these moral and legal issues is the need to solve a difficult technical problem: how to ensure the 'traceability' of RAIs in order to be able to assign moral or legal responsibility in relation to them. Traceability involves being able to determine the causes that led a RAI to behave in the way that it did, and securing

it seems especially difficult in the case of RAIs that involve bottom-up algorithms. Given the potential threat that RAIs pose to human interests, it seems hard to justify permitting their creation and use in any given case in the absence of an adequate answer to this challenge.

4.4 SIDE-EFFECTS

Side-effects, both good and bad, will inevitably result from the use of RAIs. One positive side-effect, for example, is that of affording people greater opportunities to develop their personal relationships or to pursue leisure activities. On the other hand, RAIs may cause significant levels of unemployment and, ultimately, serious levels of social deprivation, inequality, and unrest. These are, properly speaking, side-effects, since even RAIs intended to replace human labour do not have as part of their operational goal to cause unemployment or its attendant societal problems. These are at most foreseen but unintended consequences.

Economists disagree about the potential impact of the widespread use of RAIs on human employment. One study has found that up to 50% of American jobs, including jobs performed by lawyers, doctors and accountants, are at risk of being automated (Frey and Osborne 2013). In the United Kingdom, more than one in three jobs could be taken over by RAIs in the next twenty years, with the impact disproportionately affecting those in repetitive, low-paid jobs (Tovey, 2014). But these developments may not necessarily cause significant unemployment, as the history of technological innovation reveals: new jobs, many of which we currently have no inkling, might emerge, partly as a result of productivity gains made by novel technology, and often responding to new wants that technological advances have themselves helped generate (Milanovic 2016; Autor 2015). Others take a more pessimistic view, especially if they anticipate that RAIs will acquire superhuman capacities that will render pretty much all human workers obsolete (Brynjolfson and McAfee 2014; Drury 2018; Aeon 2018). Whatever the truth turns out to be, it is likely that over the shorter term many people will lose their jobs to RAIs and will face great difficulty in retraining for whatever new jobs—perhaps centred on skills requiring judgment, creativity or emotional intelligence—might eventually emerge.

These possibilities force us to reassess the value of work in human life. Partly this value has to do with earning income, and so one popular response to the danger of RAI-induced unemployment is the policy of a Universal Basic Income (UBI) fi-

nanced by the increased productivity achieved through the deployment of RAIs (Van Parijs and Vanderborght 2017).[17] A UBI would provide a regular and unconditional cash payment to all irrespective of whether they work or fulfil any further conditions such as having a disability or actively seeking work. But work also potentially serves values aside from the generation of income: it is an important source of accomplishment and self-worth, it fosters virtues of responsibility and self-discipline, and it provides a focus for valuable social engagement. Will we need to limit, in some way, the incursion of RAIs into our economic life in order to preserve adequate human access to these values? Perhaps we should use RAIs mainly as tools to assist, rather than replace, human workers. Would limiting the use of RAIs to preserve a sphere for human achievement be self-defeating, insofar as these 'achievements' would be shadowed by the knowledge that RAIs could have done an equivalent, or even far superior, job? Or are there feasible, and perhaps even preferable, ways of realizing these values through other pursuits, such as family life, art, religious worship or sport? Or, in a world liberated from the necessity of human work, might we discover an altogether new set of values to give meaning to our lives?

Unemployment is just one worrying potential side-effect of RAIs. A more diffuse worry, along broadly similar lines, is that over-reliance on RAIs may lead to the atrophying of valuable skills and to the diminution of our sense of responsibility for our own lives and choices. Doctors, drivers and pilots, for example, may begin to lose the skills that they need in order to perform well in emergencies; ordinary people may become excessively reliant on RAIs for guidance in making day-to-day decisions about such matters as the food they eat, the newspapers they read, and the political parties they vote for at elections. Moreover, the greater our reliance on RAIs in making decisions, the more we may discover that our lives are increasingly shaped by the sorts of considerations to which automated decision-making is most sensitive. This need not mean that the 'outcomes' produced are inferior to those that would have resulted had a wider range of considerations been in view, but it does mean that the contours of our lives will increasingly be a function of the capabilities of RAIs, rather than deliberation about the full array of pathways made eligible by the salient value considerations.[18]

There are other serious worries, including potential corrosive effects on the

17. Other possible responses include the provision of wage subsidies and guaranteed government employment; for a discussion of these three strategies, Furman and Seamans 2018, 21-25.

18. For a discussion of this sort of side-effect, in relation to AI encouraging a move from more 'equitable' to more 'codified' forms of legal adjudication, see Re and Solow-Niederman (2019).

quality of our relations with other humans. The more that our lives revolve around interactions with machines—often endowed with anthropomorphic forms and voices—that are fashioned to service our desires, the more we risk succumbing to the temptation to extend the same instrumentalizing attitude towards our fellow humans. This is a concern that is especially acute in relation to RAIs used for companionship or sex. Others, as we have seen, contend that we should welcome the eventual erasure of the 'human-machine divide', a process that will be hastened by the appearance of human-machine hybrids or cyborgs.

4.5 THREATS

RAIs that are specifically designed to carry out malign goals, such as privacy-violating surveillance, financial fraud or terrorist attacks, can pose serious threats to our interests and values. Such RAIs are not functional, in the sense identified above (4.1). But threats can also arise from RAIs designed to perform worthwhile tasks being sabotaged or subverted in their functioning—for example, if their algorithms are purposefully fed with false or corrupted data, or if they are hacked by malevolent agents. A world in which your smart phone spies on you or AWSs fall into the hands of terrorists is hardly a far-fetched prospect. And the threats come not only from criminals, terrorist groups or corporations, but perhaps above all from governments, often working in collaboration with groups of other kinds. A striking recent example of RAIs being used as instruments of authoritarian rule is the Social Credit System set up by the Chinese government whereby individuals receive a 'citizen score', based on data collected about them, which is used to determine eligibility for jobs, foreign travel and other benefits (*The Economist* 2016).

The response that this kind of governmental threat does not really apply to democracies is unconvincing. It grossly underestimates the ways in which politicians and bureaucrats in democracies might seize on widespread fears, for example, about terrorism or immigration, to enact repressive measures and prolong their hold on power. Another major concern is the rise of 'preventive' or 'actuarial' approaches to policing whereby AI is used to predict future crimes, and would-be criminals are apprehended on the basis of what they are likely to have done rather than any crime they have actually committed, with all the attendant risks to civil liberties vividly portrayed in the film *Minority Report* (see, for example, Ferguson 2017). Moreover, an especially insidious phenomenon in this context is the way in which some suspect

forms of government surveillance and corporate data-gathering are intertwined, as was dramatically shown by Edward Snowden's revelations about the NSA's access to data gathered by Google, Facebook, and Microsoft.[19]

One of the gravest threats posed by RAIs is to the proper functioning of democracy itself. [20] Democracy not only requires that all citizens have a vote, but also that they are able to exercise their vote after free and informed deliberation and debate on the issues that are at stake. It requires a free flow of information to enable democratic deliberation to shape policy formation and to ensure that office-holders are held accountable. Concerns have arisen in recent years about the use of RAIs to compromise these democratic processes. Methods used to this effect have included the micro-targeting of individuals with bespoke political advertisements based on data illicitly gleaned from social media platforms, such as Facebook, or using robot accounts — 'bots'—that masquerade as human in order to saturate Twitter and other platforms with propaganda, or the creation of virtually undetectable audio-visual forgeries. The threat here is not simply a matter of the illicit provenance of the data, nor even that the messages and images crafted may be deceptive or manipulative, but also the way in which such activities contribute to the formation of distinct informational 'universes' for different categories of voters, thereby eroding the common public sphere that is vital to democratic deliberation (Bartlett 2018). As to the magnitude of the threat to democracy, Onora O'Neill, has issued a stark warning:

> *Not deceiving is one of the fundamental duties. When I think about technology, I wonder whether we will have democracy in 20 years because if we cannot find ways to solve this problem, we will not. People are receiving messages and content which is distributed by robots, not by other human beings, let alone by other fellow citizens. It is frightening (O'Neill 2018; see also Helbing et al 2017).*

When allied to the potential side-effects of RAIs—undercutting both our sense of personal responsibility to our fellow citizens and our sense of the distinctiveness of humanity—the risks for democracy appear significant.

Ways of addressing the threat to democracy include technology-specific mea-

19. For a discussion, including the point that government agencies often circumvent legal restrictions on data-gathering by buying, demanding, or hacking data held by corporate agents, see Pasquale 2015, 48-51.

20. For an overview of the implications, both positive and negative, of digital technology for democracy, see Susskind (2018).

sures such as enhanced privacy protection for personal data, greater transparency regarding the use of data by online platform providers, and more stringent registration processes for social media accounts. O'Neill even raises the possibility of some form of internet censorship akin to that practiced by the Chinese government, which may be a cure worse than the disease it aims to treat. But it is important also to address structural features of our political systems that may interact with RAIs to subvert democracy. For example, the United States' lax campaign-financing laws make it easier for resources to be funnelled into the large-scale dissemination of 'fake news'. More generally, we face a conundrum. What is arguably needed to counter most of the societal threats posed by RAIs is enhanced democratic accountability; but this may involve us in a race against time, since one of the gravest of those threats is precisely to the functioning of democratic processes. We need democratic solutions to the problems posed by RAIs before they are used to destroy democracy itself.

Of course, according to some, the greatest threat posed by RAIs is the one from which this article began: that they become so much more intelligent than humans that they eventually subjugate or eradicate us in pursuit of their own ends. This doomsday scenario has been emphasized by prominent figures in the field of RAIs, such as Bill Gates and Elon Musk, as well as by leading scientists, including the late Stephen Hawking, who observed towards the end of his life:

> *A super-intelligent AI will be extremely good at accomplishing its goals, and if those goals aren't aligned with ours, we're in trouble... You're probably not an evil ant-hater who steps on ants out of malice, but if you're in charge of a hydroelectric green energy project and there's an anthill in the region to be flooded, too bad for the ants. Let's not place humanity in the position of those ants (Griffin 2015).*

Some have dismissed such warnings as irresponsibly speculative on the basis that general AI, let alone general AI of a superhuman form, is not a realistic prospect in the foreseeable future. As Daniel Dennett (2019) has put it, 'we're making tools not colleagues'. On this view, doomsday scenarios are fantasies that distract us from other, urgent RAI-related problems that we confront. But another response focuses on the assumption that underlies Hawking's warning. Why assume that the goals of super-intelligent RAIs will be troublingly unaligned with ours? If RAIs develop truly superhuman abilities, won't these include the abilities to reason about, and

conform to, morality?[21] In other words, it is an impoverished concept of 'intelligence' to confine it to the capacity to achieve complex goals, regardless of their moral value. Imagine, then, a world in which we are governed by just and benevolent RAIs that far surpass any human in intelligence and goodness. Is this scenario the ultimate fulfilment of the promise of RAIs to serve humankind or a deep betrayal of our interests and values? That there is a case for the latter conclusion is evident. Partly, this is a matter of being subjected to forms of governance based on principles and considerations that potentially outrun the ability of most, or perhaps all, human beings to grasp, which is a massive deviation from the Enlightenment idea of rule under standards that can be rationally apprehended and approved by those subjected to them.[22] After all, given that human morality is attuned to the possibilities and limitations inherent in our human predicament, why should it be supposed that super-intelligent RAIs, who do not share a human nature, would be disposed to give much weight to anything we could recognize as moral considerations? But leaving this worry aside, a key value that the putative just and benevolent RAI governors would presumably have to acknowledge is that of human freedom, not only at the level of the individual human making personal life choices, but at the level of groups of human beings exercising their communal self-determination. It is difficult to see how their rule could avoid seriously undermining this freedom.[23] Perhaps, in light of these considerations, a benign race of super-RAIs would turn out to have the grace to leave humans to forge their own path, subject to the enforcement of a minimal set of norms that averted the more devastating manifestations of human error.

5. CONCLUSION

The task of developing a sound approach to the ethics of RAIs operates at multiple levels, including legal regulation, social morality, and personal moral standards, which interact in complex ways. The main first-order questions arising at these three

21. For a more elaborate defence of the opposite, non-alignment thesis, as to the values of RAIs, see Bostrom (2014).

22. A point emphasized in Kissinger (2018) 'The most difficult yet important question about the world into which we are headed is this: What will become of human consciousness if its own explanatory power is surpassed by AI, and societies are no longer able to interpret the world they inhabit in terms that are meaningful to them?'

23. A further, more speculative consideration here is that rule by non-humans would be an attack on loyalty to, or identifying with, one's species, see Williams 2006, 149-152.

levels, I have suggested, can be for the most part illuminatingly collected under the FIRST schema, as pertaining to the rubrics of functionality, inherent significance, rights and responsibilities, side-effects, and threats. However, I have cast doubt on the idea that there is some helpful segment of existing moral or legal principle that primarily or exclusively bears on the ethics of RAIs. Instead, RAIs should be seen as potentially engaging the full gamut of human values under the five rubrics identified by the FIRST schema. Although work has already been done under all five of these rubrics, we are at an early stage of thinking through the ethics of RAIs. Moreover, the inherent significance of the human factor is a matter that has not as yet, for understandable reasons, received anything like the level of sustained attention it deserves. We need to grapple with the very idea of such significance and how it acquires varying forms of moral valence, depending upon features such as the domain of decision-making (e.g. cancer diagnosis or criminal sentencing) or the form of relationship (e.g. lawyer or lover) that is at issue. In addressing questions that arise under these five rubrics, it is essential to be guided and constrained by a realistic appreciation of the existing and foreseeable future capacities of RAIs, and not to allow our ethical thought to be hi-jacked by utopian (or dystopian) speculation based on possibilities that lie, at best, in the remote future, even if they do not strictly fall outside the realm of scientific and technological possibility.

ACKNOWLEDGEMENTS

I am grateful to Roger Brownsword, Rebecca Lowe, Claudia Chwalisz, Francesca Rossi, Jose Such, David Nelken, participants at a KCL/Queen's Ontario colloquium in legal philosophy, and—especially—Hannah Maslen, Annette Zimmermann, and an anonymous referee for helpful comments on previous drafts. Work on this paper was enabled by a grant from the Future of Life Institute. I am also grateful for the excellent research assistance provided by Napoleon Xanthoulis. A condensed version of this article is published as Tasioulas (forthcoming).

REFERENCES

Aeon. (2018). "Humans Need Not Apply." (video) *Aeon*, <https://aeon.co/videos/the-robots-are-coming-for-our-jobs-why-the-human-workforce-is-at-risk> [Accessed on 13 June 2019].

Asilomar AI Principles. (2017). <https://futureoflife.org/ai-principles/> [Accessed on 13 June 2019].

Asimov, I. (1950). "Runaround.". I, *Robot*. New York City: Doubleday.

Autor, D. (2015). "Why are There Still So Many Jobs? The History and Future of Workplace Automation." *Journal of Economic Perspectives* 29 (3): 30.

Barocas, S. and Selbst, A. (2016). "Big Data's Disparate Impact." *California Law Review* 104: 670-732.

Bartlett, J. (2018). The *People vs Tech: How the Internet is Killing Democracy (and How We Save It)*. London: Penguin.

Binns, R. (2018). "Fairness in Machine Learning: Lessons from Political Philosophy." *Journal of Machine Learning Research* 81: 1-11 <https://arxiv.org/abs/1712.03586> [Accessed on 13 June 2019].

Bostrom, N. (2014). *Superintelligence: Paths, Dangers, Strategies*. Oxford: Oxford University Press.

Bonnefon, J.-F., Shariff, A., and Rahwan, I. (2006). "The Social Dilemma of Autonomous Vehicles." *Science* 352: 1573-1576. <http://science.sciencemag.org/content/352/6293/1573> [Accessed on 13 June 2019].

Bryson, J, Diamantis, M. and Grant, T. (2017). "Of, for and by the People: The Legal Lacuna of Synthetic Persons." *Artificial Intelligence and Law* 25: 273-29.

Buolamwini, J. and Gebru, T. (2018). "Gender Shades: Intersectional Accuracy Disparity in Commercial Gender Classification." *Proceedings of Machine Learning Research* 81: 1-15. <http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf> [Accessed on 13 June 2019].

Buranyi, S. (2018). "Dehumanising, impenetrable, and frustrating: the grim reality of job-hunting in the age of AI." *Guardian* March 4th <https://www.theguardian.com/inequality/2018/mar/04/dehumanising-impenetrable-frustrating-the-grim-reality-of-job-hunting-in-the-age-of-ai> [Accessed on 13 June 2019].

Brynjolfson, E. and McAfee, A. (2014). *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York City, NY: WW Norton & Co.

Burgess, M. (2017). "Just Like Humans, Artificial Intelligence Can Be Sexist and Racist." *Wired* April 13. <http://www.wired.co.uk/article/machine-learning-bias-prejudice> [Accessed on 13 June 2019].

Coeckelbergh, M. (2015). "Artificial Agents, Good Care, and Modernity." *Theoretical Medicine and Bioethics* 36: 265-277.

Danaher, J. and McArthur, N. Eds. (2017), *Robot Sex: Social and Ethical Implications*. Cambridge MA: MIT Press.

Dennett, D.C. (2010). "Will AI Achieve Consciousness? Wrong Question." *Wired* (Feb 9) <https://www.wired.com/story/will-ai-achieve-consciousness-wrong-question/> [Accessed on 13 June 2019]

Drury, C. (2018). "Mark Carney Warns Robots Taking Jobs Could Lead to Rise of Marxism." *The Independent* April 14 <https://www.independent.co.uk/news/uk/home-news/mark-carney-marxism-automation-bank-of-england-governor-job-losses-capitalism-a8304706.html> [Accessed on 13 June 2019].

Edmonds, D. (2017). "Can We Teach Robots Ethics?" *BBC News*, October 15, 2017 <http://www.bbc.co.uk/news/magazine-41504285> [Accessed on 13 June 2019].

European Commission (2019). *Ethics Guidelines for Trustworthy AI*. European Commission: Brussels. <https://ec.europa.eu/futurium/en/ai-alliance-consultation> [Accessed on 13 June 2019].

European Parliament (2017). Report with recommendations to the Commission on Civil Law Rules on Robotics, Jan 27. <http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-//EP//NONSGML+TA+P8-TA-2017-0051+0+DOC+PDF+V0//EN> [Accessed on 13 June 2019].

Ferguson, A.G. (2017). *The Rise of Big Data Policing*. New York City, NY: New York University Press.

Frey, C.B. and Osborne, M.A. (2013). "The Future of Employment: How Susceptible Are Jobs to Computerisation." <https://www.oxfordmartin.ox.ac.uk/downloads/academic/The_Future_of_Employment.pdf> [Accessed on 13 June 2019].

Fry, H. (2018). *Hello World: How to be Human in the Age of the Machine*. New York: Doubleday.

Furman, J. and Seamans, R. (2018). "AI and the Economy." <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3186591> [Accessed on 13 June 2019].

Goode, L. (2018). "Facial Recognition Software is Biased Towards White Men, Researcher Finds." *The Verge* (Feb 11). <https://www.theverge.com/2018/2/11/17001218/facial-recognition-software-accuracy-technology-mit-white-men-black-women-error> [Accessed on 13 June 2019].

Griffin, A. (2015). "Stephen Hawking: Artificial Intelligence Could Wipe Out Humanity When It Gets Too Clever as Humans Will be Like Ants." *The Independent* (October 8) <https://www.independent.co.uk/life-style/gadgets-and-tech/news/stephen-hawking-artificial-intelligence-could-wipe-out-humanity-when-it-gets-too-clever-as-humans-a6686496.html> [Accessed on 13 June 2019].

Gunkel, D. (2018). *Robot Rights*. Cambridge, MA: MIT Press.

Helbing, D., Frey, B.S., Gigerenzer, G., Hafen, E., Hagner, M., Hofstetter, Y., van den Hoven, J., Zicari, R.V., and Zwitter, A. (2017). "Will Democracy Survive Big Data and Artificial Intelligence?" *Scientific American* (Feb 25 <https://www.scientificamerican.com/article/will-democracy-survive-big-data-and-artificial-intelligence/?redirect=1> [Accessed on 13 June 2019].

House of Lords AI Committee (2018). *AI in the UK; ready, willing and able?* House of Lords Paper 100. <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm> [Accessed on 13 June 2019].

Jordan, M. (2018). "Artificial Intelligence—The Revolution Hasn't Happened Yet." *Medium* April 18, <https://medium.com/@mijordan3/artificial-intelligence-the-revolution-hasnt-happened-yet-5e1d5812e1e7> [Accessed on 13 June 2019].

Kissinger, H.A. (2018). "How the Enlightenment Ends." *The Atlantic* (June) <https://www.theatlantic.com/magazine/archive/2018/06/henry-kissinger-ai-could-mean-the-end-of-human-history/559124/> [Accessed on 13 June 2019].

Kleinberg, J., Ludwig, J., Mullainathan, S., and Sunstein, C. (forthcoming). "Discrimination in the Age of Algorithms." <https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3329669> [Accessed on 13 June 2019].

Kurzweil, R. (2002). "We Are Becoming Cyborgs." <http://www.kurzweilai.net/we-are-becoming-cyborgs> [Accessed on 13 June 2019].

——— (2005). *The Singularity is Near: When Humans Transcend Biology*. London: Duckworth.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. (2016). "How We Analysed the COMPAS Recidivism Algorithm." *ProPublica* May 23 <https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm> [Accessed on 13 June 2019].

Milanovic, B. (2016). "Three Fallacies that Make You Fear a Robot Economy." *Evonomics*, Sept 1. <http://evonomics.com/three-fallacies-robot-economy-branko/> [Accessed on 13 June 2019].

Nemitz, P. (2018). "Constitutional Democracy and Technology in the Age of Artificial Intelligence." *Philosophical Transactions Royal Society A* 376 (2133).

Nyholm, S. (2018). "The Ethics of Crashes with Self-Driving Cars: A Road Map, I." *Philosophy Compass* 17.

O'Neil, C. (2016). *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*. London: Penguin.

———(2018). "I Wonder Whether We Will Have Democracy in 20 Years." Interview with Elena Cué. *Huffington Post* May 1st. <https://www.huffingtonpost.com/entry/onora-oneill-i-wonder-whether-we-will-have-democracy_us_5a4f8a12e4b0cd114bdb324f> [Accessed on 13 June 2019].

Oswald M., Grace, J., Urwin, S., and Barnes, G.C. (2018). "Algorithmic Risk Assessment Policing Models: Lessons from the Durham HART Model and 'Experimental' Proportionality." *Information & Communications Technology Law* 27 (2018): 223.

Pasquale, F. (2015). *The Black Box Society: The Secret Algorithms that Control Money and Information*. Cambridge MA: Harvard University Press, 2015.

Pessoa, L. (2018). "Robot cognition requires machines that both think and feel." *Aeon* April 13. <https://aeon.co/ideas/robot-cognition-requires-machines-that-both-think-and-feel> [Accessed on 13 June 2019].

Re, R.M. and Solow-Niederman, A. (forthcoming). "Developing Artificially Intelligent Justice." *Stanford Technology Law Review* 22.

Rini, R. (2017). "Raising Good Robots." *Aeon* April 18 <https://aeon.co/essays/creating-robots-capable-of-moral-reasoning-is-like-parenting> [Accessed on 13 June 2019].

Savulescu, J. and Maslen, H. (2015). "Moral enhancement and artificial intelligence: moral AI?" *Beyond Artificial Intelligence: The Disappearing Human-Machine Divide*. Eds. Romportl, J., Zackova, E., and Kelemen, J. Switzerland: Springer International Publishing.

Schwitzgebel, E. and M. Garza, M. (2015). "A Defense of the Rights of Artificial Intelligences." *Midwest Studies in Philosophy* 39.

Sparrow, R. (2007). "Killer Robots." *Journal of Applied Philosophy* 24: 62-77.

Sunstein, C. (forthcoming). "Algorithms, Correcting Biases." *Social Research*.

Susskind, J. (2018). *Future Politics: Living Together in a World Transformed by Tech*. Oxford: Oxford University Press.

Tasioulas, J. (2019). "AI and Robot Ethics." *Ethics and the Contemporary World*. Ed. Edmonds, D. London: Routledge.

———(forthcoming). "Saving Human Rights from Human Rights Law." *Vanderbilt Journal of Transnational Law*.

Tegmark, M. (2017). *Life 3.0: Being Human in the Age of Artificial Intelligence*. London: Penguin.

The Economist. (2016). "Big Data, Meet Big Brother: China Invents the Digital Totalitarian State." *The Economist* December 17. <https://www.economist.com/briefing/2016/12/17/china-invents-the-digital-totalitarian-state> [Accessed on 13 June 2019].

Tovey, A. (2014). "Ten Million Jobs at Risk from Advancing Technology." *Telegraph* (Nov. 10) <https://www.telegraph.co.uk/finance/newsbysector/industry/11219688/Ten-million-jobs-at-risk-from-advancing-technology.html> [Accessed on 13 June 2019].

Turner, J. (2018). *Robot Rules*. London: Palgrave MacMillan.

UNESCO (2017). Report of COMEST on Robotics Ethics. Paris, UNESCO. <http://unesdoc.unesco.org/images/0025/002539/253952E.pdf> [Accessed on 13 June 2019].

Upchurch, T. (2018). "To Work for Society, Data Scientists Need a Hippocratic Oath With Teeth." *Wired*, April 8th. <http://www.wired.co.uk/article/data-ai-ethics-hippocratic-oath-cathy-o-neil-weapons-of-math-destruction> [Accessed on 13 June 2019].

Van Parijs, P. and Vanderborght, Y. (2017). *Basic Income: A Radical proposal for a Free Society and a Sane Economy*. Cambridge, MA: Harvard University Press.

Vayena, E. and Tasioulas, J. (2016). "The Dynamics of Big Data and Human Rights: The Case of Scientific Research." *Philosophical Transactions of the Royal Society A* 28: 374.

Wachter, S. Mittelstadt, B., and Floridi, L. (2017). "Why a Right to an Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation." *International Data Privacy Law 7*: 76-99.

Wiggins, D. (2016). "Sameness, Substance, and the Human Person." *Continuants: Their Activity, Their Being and Their Identity—Twelve Essays*. Oxford: Oxford Unversity Press.

Williams, B. (2006). "The Human Prejudice." *Philosophy as a Humanistic Discipline*. Princeton: Princeton University Press.