The Economics of Morality

OXFORD UEHIRO PRIZE IN PRACTICAL ETHICS 2014-15

DILLON BOWEN

Tufts University

ABSTRACT

Altruism is embedded in our biology and in our culture. We offer our bus seats to the disabled and elderly, give directions to disoriented tourists, and donate a portion of our income charity. Yet for all the good it does, there are deep problems with altruism as it is practiced today. Nearly all of us, when asked, will say that we care about practicing altruism in a way that effectively improves the lives of others. Almost none of us, when asked, can honestly say that we have made a serious effort to ensure that we are practicing altruism in a way that effectively improves the lives of others. Disparities like these are indicative of flaws in our cognitive architecture - biases which ensure that the traditional practice of altruism is incongruous with our own values. This disconnect between our values and our actions causes our altruistic efforts to help fewer people to a lesser extent than they otherwise could. I argue that traditional altruism is in need of reformation and defend a social and philosophical movement aimed at achieving this reformation known as effective altruism. The reason effective altruism is such a promising alternative to traditional altruism is its application of economic thinking to the realm of altruism and morality. An economist's mentality is, I suggest, a necessary instrument for bridging the gap between our values and our actions, allowing us to practice altruism in a way that more effectively improves the lives of others.

INTRODUCTION

People perform acts of altruism every day. When I describe an act as *altruistic*, I mean that the person performing the act (the *donor*) makes a personal sacrifice—perhaps in terms of time or money—for the sake of improving the well-being of another conscious creature (the *recipient*). In this context, we will find it helpful to narrow the definition of *altruism* to describe only those altruistic actions in which the recipient is not a member of the donor's family, friends, or community. For the purposes of this paper, an action can be altruistic only if the donor has little expectation that she will have a personal or economic relationship with the recipient. Altruism can be anything from holding the door for a stranger to donating a substantial amount of money to charity. Almost everyone, I wager, behaves altruistically from time to time —some of us on a daily basis.

The problem with altruism, as it is currently practiced, is that it is ineffective at improving the lives of conscious creatures. In what sense is the ineffectiveness of altruism 'problematic'? Instead of appealing to moral obligations or duties, I will argue that the ineffectiveness of altruism is problematic in the sense that most of those who practice altruism would, on reflection, prefer to do so more effectively. When we behave as *ineffective altruists*, we are therefore failing to behave in accordance with our own preferences. The alternative is an ethical framework known as *effective altruism*, which is most concisely described as "aiming to do the most good that one can". (Singer & MacAskill 2015, p.viii)

This paper is divided into four sections. The first gives a more rigorous definition and explanation of effective altruism. Following this, I explore the implications of effective altruism for population ethics, and show it to be a milder and more intuitive philosophy than its close cousin, classical utilitarianism. The third section explains how cognitive biases cause us to behave as ineffective altruists, and suggests that our preferences would be better served by practicing altruism more effectively. Finally, I draw an analogy between how we think about altruism and how we think about economics. As I hope to show, thinking of altruism economically will aid us in overcoming the cognitive biases that make altruism so ineffective.

EFFECTIVE ALTRUISM

Singer & MacAskill 2015 describes effective altruism as "aiming to do the most good that one can". (ibid) While this definition succeeds in its concision and popular appeal, it leaves something to be desired in terms of specificity. We might wonder, for example, what is meant by *doing good*, and if there are any bounds on the amount of time and money effective altruists should devote to doing good. I offer my own definition here in hopes that it will help clarify some of these questions.

Effective altruism is the belief that we should endeavour to spend whatever resources we plan to devote to valuable creatures who are unlikely to have a substantial impact on our lives in such a way as to maximize their aggregate well-being, provided we do not sacrifice anything else of importance in doing so.

Suppose I plan to donate $100 to charity, and that for some reason I have to choose between two charities—A and B. Both A and B provide deworming treatments for people in Kenya. For the same $100, A can deworm two people, but B can deworm only one. Assuming A and B have similar externalities, I ought to donate to the charity which provides deworming treatments for two people rather than one. All else being equal, effective altruism holds that we should improve the lives of as many people as we possibly can. Call this the *helping more people (HMP)* imperative.

Now imagine I am faced with a different choice of charities—C and D. Both C and D feed families in Uganda. For the same $100, C can feed a family for two months, but D can feed a family for only one month. Assuming C and D have similar externalities, I ought to donate to the charity which feeds a family for a longer period of time. All else being equal, effective altruism holds that we should improve people's lives to the greatest extent we can. Call this the *helping people more (HPM)* imperative.

Presented this way, effective altruism seems like a straightforward and appealing ethical philosophy. These are, of course, the easy cases. To think about more difficult cases, it will help to examine each piece of my definition in turn.

We should endeavour to spend whatever resources we plan to devote…

My view of effective altruism is weaker than what Singer wants to propose. Singer has argued, in previous works, that we should devote as much of our time and money to others as possible, stopping only when the marginal utility of keeping money for ourselves outweighs the marginal utility of donating money to others. (see, e.g. Singer 1972) Though I strongly believe we ought to devote more of our time and money to helping others than we currently do, all I want to claim here is that whatever resources we would have spent helping others in any case should be spent in such a way as to maximize the aggregate well-being of *valuable creatures*.

...to valuable creatures...

Who is included in the set of creatures whose aggregate well-being we are trying to maximize? In other words, who should be the recipients of our altruism? I designate a set which I call *valuable creatures*. Who exactly is included in this set may depend on the donor's preferences. For example, a classical utilitarian would consider all beings capable of experiencing happiness and suffering—both those that currently exist and all those that could potentially exist in the future—as morally important. An anti-natalist, by contrast, values only creatures who currently exist and whose birth cannot be prevented. We may, as I do, wish to include nonhuman animals and artificial intelligences in this set, or we may not. I leave this category purposefully vague.

In order to avoid repeating the awkward verbiage *valuable creatures*, I will often refer to the recipients of our altruism simply as *people*. Please understand that this term is not meant to exclude nonhuman creatures.

...who are unlikely to have a substantial impact on our lives...

To reiterate a point I made in the introduction, the altruistic actions under consideration here are those in which the donor does not expect to have a personal or economic relationship with the recipient.

...in such a way as to maximize their aggregate well-being...

When evaluating the impact of an altruistic action, effective altruists care about 1) how many people it helps (HMP imperative) and 2) how much it helps them (HPM imperative). But what happens when these measures come into conflict? For example, imagine I have to choose between charities E and F, both of which fight malaria by providing long lasting insecticidal bed-nets to villages in Malawi. Charity E will use my $100 donation to provide bed-nets for two villages for one year. Charity F will use my $100 donation to provide bed-nets for one village for two years. E helps more people, but F helps people more. Assuming the externalities of both these charities are the same, which should an effective altruist donate to?

To address questions like these, I collapse the measures of the HMP and the HPM imperatives into a single scale—aggregate well-being. If there are no further considerations that would weigh in favour of charities E or F, an effective altruist should be indifferent between them.

...provided we do not sacrifice anything else of importance in doing so.

However, we might think that there *are* further considerations which would allow us to choose between E and F. One could argue that E is a fairer charity, because it increases living standards of as many at-risk communities as it can. F is behaving unfairly, the argument goes, in providing a single village with two years of security given that children in surrounding villages are dying from malaria every day. If a particular donor, compelled by this line of reasoning, chose E over F, would that disqualify her as an effective altruist?

No. This is the purpose of the final clause of my definition. Most people profess to hold values which are not reducible to measures of well-being. If there are other important considerations weighing against aggregate well-being, it may be rational for a donor to prefer one altruistic action over another, despite them being equally effective. It may even be rational for a donor to prefer a less effective altruistic action over one which is more effective if these considerations are sufficiently compelling.

I hasten to clarify that this clause is meant to make room for ineffective altruism only when it is based on what a donor would rationally endorse as an *important* consideration. For instance, men tend to donate more generously to a charity when solicited by an attractive female. (Raihani & Smith 2015) Presumably the gender and aesthetic appeal of a charity solicitor does not qualify as an important consideration for most people, and therefore donating to an ineffective charity on this basis would be out of keeping with effective altruism.

Now that we have gone into some detail about what effective altruism is, we can discuss its implications for difficult cases—specifically its implications for two of population ethics' most obstinate problems—the repugnant conclusion and the non-identity problem. In doing this, I intend to show the plausibility of effective altruism in even the thorniest of philosophical issues, and distinguish it from its counterintuitive cousin, classical utilitarianism. Most of the theoretical objections I have encountered to effective altruism centre around its ostensibly objectionable stance on population ethics, so it is important to set the record straight on this matter before moving on to pragmatic considerations.

EFFECTIVE ALTRUISM AND POPULATION ETHICS

The repugnant conclusion

One argument I frequently encounter against effective altruism runs like this:

If I accept effective altruism, I must accept the repugnant conclusion

I reject the repugnant conclusion

Therefore, I reject effective altruism

Just what is the repugnant conclusion, and why might we believe that effective altruism entails it? The repugnant conclusion (Parfit 1984) was first raised as an objection to classical utilitarianism, which holds that the one and only good is to maximize *aggregate* well-being. The objection attempts to invalidate classical utilitarianism on the grounds that it concerns itself solely with aggregate well-being and ignores *average* well-being. To see why we might desire an ethical philosophy that concerns itself with average well-being, imagine three worlds—A, B, and C. World A is home to only a few people (say, 10 people, or n=10), all of whom are extremely happy (whose level of well-being is 10, or u=10). By contrast, world B is home to very many people (n=100) whose lives are barely worth living (u=1). The inhabitants of world C also have lives that are barely worth living (u=1), but there are more of them than in world B (n=101). According to classical utilitarianism, we should be indifferent between worlds A and B (total utility=100), and prefer C to both of them (total utility=101). The repugnant conclusion is that, for any given world, a classical utilitarian will always prefer a world full of people whose lives are just barely worth living, so long as there are enough of them to offset the decrease in average happiness. Surely, the argument goes, we must reject the repugnant conclusion, and therefore classical utilitarianism.

Effective altruism is similar to classical utilitarianism in that it advocates maximizing the aggregate well-being of valuable creatures. In fact, classical utilitarianism is a form of effective altruism. The concern is that, by focusing only on aggregate well-being to the exclusion of average well-being, effective altruism makes the same mistake classical utilitarianism does. However, there is an important difference between effective altruism and utilitarianism which makes effective altruism compatible with a rejection of the repugnant conclusion.

Recall that effective altruism holds that we should maximize the well-being of 'valuable creatures', while being purposefully vague about which creatures are included in this set. A classical utilitarian has a precise view of which creatures are morally important—all of them, including all creatures alive today and which may potentially exist in the future. Even if a classical utilitarian would prefer to prevent someone from being born—say, a child who would have a debilitating illness with a high mortality rate in the first years of life—she would still consider this child a valuable creature. If it were possible, the classical utilitarian would rather see this child born and live a happy, healthy life.

But effective altruists are not committed to adopting such a broad set of valuable creatures. Take, for example, average utilitarianism, which holds that the one and only good is to maximize the average well-being of existing creatures. To the average utilitarian, the set of valuable creatures consists of those who will have a positive impact on average utility and those whose existence cannot be terminated or prevented without diminishing average utility by an even greater amount. Imagine we live in a world with a few (n=10) very happy people (u=10). Further imagine that one couple is considering having a child whose life, for whatever reason, will barely be worth living (u=1). An average utilitarian would prefer that this couple refrained from having a child.

By contrast, a classical utilitarian would prefer the couple *did* have their child. After all, it will increase aggregate utility, if only by a small increment. The difference of opinion between average and classical utilitarianism results from how they view the set of valuable creatures. Since the child's life diminishes average well-being, the average utilitarian considers it morally important if and only if its existence cannot be prevented without an even greater decrease to average utility. But the classical utilitarian views the potential child as morally important whether or not it is actually born. Whereas the average utilitarian would view the couple's choice not to have the child as excluding it from the set of valuable creatures, the classical utilitarian would view this choice as diminishing the child's utility from small to zero.

Is average utilitarianism a version of effective altruism? Yes it is. For any finite set of valuable creatures, average utility is maximized when aggregate utility is maximized. Does average utilitarianism avoid the repugnant conclusion? Again, the answer is yes. An average utilitarian would prefer world B (n=10, u=10) to worlds A (n=100, u=1) and C (n=101, u=1), and in all cases prefers a world with fewer, happier people to a world with more people whose lives are barely worth living.

Bringing the discussion back to the larger picture, I should add that I am not an advocate of average utilitarianism, which yields many counter-intuitive conclusions of its own. I bring up this ethical view because it is an example of how we can be effective altruists and still reject the repugnant conclusion. Furthermore, we can see that the way to do this is by limiting the set of valuable creatures. Effective altruism entails the repugnant conclusion if and only if we consider all people currently alive and all people with the potential to be born morally important. But such a position is not logically entailed by effective altruism.

The non-identity problem

Another objection to effective altruism which similarly relies on population ethics considerations, relies on the non-identity problem[1]:

> *If I accept effective altruism, I must accept that I can be morally blameworthy for actions which are not bad for anyone*
>
> *I reject the idea that I can be morally blameworthy for actions which are not bad for anyone*
>
> *Therefore, I reject effective altruism*

The non-identity problem involves a conflict of intuitions. At first, it seems that an action can only be bad if it is bad *for* someone. An action that neither harms nor is in any way bad for someone seems as if it cannot be wrong. But now consider a 14-year old girl who is thinking of having a child. If she decides to go through with the pregnancy, her child would live a worthwhile life. However, given her age and socioeconomic status, she will not be able to provide as good a life for her baby as she

[1] The non-identity problem was first discussed in Parfit 1984.

would be able to if she waited until, say, age 26 to start a family. The intuition here is that getting pregnant at her age would be wrong.

But supposing the girl's own well-being is not affected, for whom would this action be wrong? The tempting answer is to say that it is wrong for her child. Yet the child she would have at age 14 would live a worthwhile life, and the child she would have at age 26 would be a fundamentally different person, having a different genetic structure and growing up in a different environment. So postponing pregnancy would not so much make life better for her child as it would change the identity of her child. In other words, the decision to wait to have a baby would not make life better for the child she would have had, but rather would create a different child who would lead a better life. Having a child at age 14, then, is not bad *for* anyone.

The same line of reasoning can apply to all future people. Many of the ways to 'improve' the lives of future people do not improve the lives of the future people who would have existed anyway, but rather create a different set of future people who would lead better lives. There may be very few ways to improve or diminish the quality of life of future people without changing their identities. Combine the fact that future people have undetermined identities with the moral principle that actions can only be good or bad if they are good or bad *for someone*, and we might conclude that the moral obligations we have to future people are highly limited.

What does this have to do with effective altruism? The idea is that most effective altruists include future people in their set of valuable creatures, and believe that our actions can be good or bad in relation to future people. But such a view contradicts the moral principle that actions can only be bad if they are bad *for someone*.

I believe the best way to respond to this objection is by referencing a point I made in the introduction to this piece. Ineffective altruism, I said, is problematic in the sense that it violates our preferences. When people behave as an ineffective altruists, I do not necessarily think they are violating a moral duty so much as behaving in a way I disapprove of, and a way they themselves would probably disapprove of in light of their own values. We could censure a 14-year old girl who decides to have a child on similar grounds. It may not be the case that she is violating a moral duty, but it is the case that we would prefer she made a different decision and, on reflection, she probably would as well.

This standard applies to considerations of future people in general. Imagine you can press either a red or blue button. The red button will determine that, a century from now, the world will be filled with extremely happy people. The blue button will determine that, in the same amount of time, the world will be filled with the same number of people whose lives are only moderately happy. Further suppose the identities of the people in both these worlds are fundamentally different. If someone chose to push the blue button, it would be entirely reasonable to conclude that she has done something bad. And what makes this action bad is not necessarily that it is bad *for someone*, but that it creates a suboptimal world as judged by our values.

There are two senses of *bad* at play here. One sense implies a violation a moral duty and thereby moral blameworthiness. The other implies a violation of our preferences, and thereby social disapprobation. I would argue for an interpretation of effective altruism in which a disregard for future people is bad in the latter sense but not necessarily the former. Effective altruism does not imply moral blameworthiness for actions which are not bad for anyone, but rather strongly suggests that, in light of our own values, we should perform actions which maximize the aggregate well-being of future as well as existing people.

Effective altruism does not logically entail counterintuitive conclusions about population ethics. We do not need to accept the repugnant conclusion or believe that we are morally blameworthy for actions which are not bad for anyone in order to be effective altruists. It is interesting to note that the philosopher who first discussed the repugnant conclusion and the non-identity problem, Derek Parfit, is one of effective altruism's most vocal proponents today. Effective altruism is a much less radical proposition than utilitarianism and, as I hope I have shown, an extremely sensible moral philosophy. However, we might wonder, if effective altruism is so intuitively and logically appealing, why is altruism today so ineffective at improving the well-being of valuable creatures?

ALTRUISM AS PRACTICED TODAY

Most People are Ineffective Altruists

Altruism can take many forms, but for this section I will focus on charitable giving. Many people act as if under the impression that all charities are equally good. But if 'equally good' is taken to mean 'equally effective at improving people's lives', the claim becomes immensely implausible. The notion that all charities are equally good at helping people is about as likely to be true as the notion that all companies are equally good at producing quality commodities. Why would it be the case that all charities currently in existence just happen to be equally effective at alleviating suffering?

Suppose we reject the belief that all charities are equally good. There is still the epistemic problem of determining which charities are better than others, and particularly, which charities are the best of them all. Those wishing to object here might claim that there are, at present, no means by which to determine how effective charities are. The claim that we have no way of knowing which charities are better than others is only slightly more plausible than the claim that no charity is, in fact, better than another. To maintain such a belief, we would have to conclude that Homeopaths Without Borders (yes, this is a real charity) is, for all we know, just as effective at improving well-being as any other charity in existence.

Here is a concrete example to illustrate the difference between effective altruism and ineffective altruism. Suppose we plan to donate $40,000 to prevent or alleviate the symptoms of blindness. Providing a single blind person with a guide dog will cost the entire $40,000 (Ord 2013 p.1) By contrast, the cost of surgery to cure trachoma-induced blindness is less than $20.(Ibid) With $40,000 one could either provide a single blind person with a guide dog, or cure 2,000 people in the developing world of trachoma-induced blindness. Conservatively estimating that the quality of life improvement of providing someone with a guide dog is equal to that of curing someone of trachoma-induced blindness, the choice is clear.

Proponents of the 'uncertainty argument' outlined above would have to believe these estimates so inaccurate as to have misassessed the situation by three orders of magnitude. Hopefully this possibility is sufficiently unlikely to compel us to accept two conclusions. First, charities differ in the degree to which they improve the lives of conscious creatures. And second, that the information needed to accurately assess cost-effectiveness is at least partially available.

If people were genuinely motivated to give to charity based on an intrinsic desire to improve the well-being of others, we might assume they would spend at least a bit of time and effort attempting to find this information. But this is not the pattern of behavior we observe. 83% of Americans donate to charity. (Gallup Editors 2013) Of them, 10% say they do not care at all about non-profit performance. (Hope Consulting 2011) The rest *say* they care about non-profit performance, but only 3% have done any research to find the highest performing charities. (Ibid)

However, you might wish to object, maximizing your impact does not necessarily require researching high-impact charities. For instance, you might think that, instead of spending an hour googling effective charities, you could spend another hour at work to earn more money to donate. This is an interesting possibility, but highly implausible. Given the amount of time people spend working and the amount of money people donate to charity, such a move would only be rational if donors expected an hour's worth of research to yield less than a 0.05% increase in the effectiveness of their giving[2]. It is also worth noting that donors who do not research never cite anything like this as their reason for not conducting research—the closest equivalent being that 4% of them say they are too lazy.(Hope Consulting 2011)

Despite professing to care about effectiveness when asked, most people practice altruism ineffectively. This means that those who claim to care about effectiveness either hold beliefs about charity which are fantastically detached from reality or are being insincere. My vote is for the latter. The cost of providing one guide dog for one blind person is the equivalent of curing 2,000 people of trachoma-induced blindness. Every dollar we donate to someone in poverty in the developed world could have been donated to someone 20 times as destitute in the developing world[3]. The money required to grant a single wish for a terminally ill child could have saved five children from dying in the first place[4]. Yet we continue to donate massive sums of money to ineffective charities, and our donations will achieve only a small fraction of their potential to reduce suffering.

[2] In 2014, US donors gave $358 billion to charity, or about 2% of annual GDP (Giving USA 2014). Adjusting for the fact that only 83% of Americans, donate, this makes 2.5% per donor on average. My calculations assume that individuals give at this rate throughout their lives. The average person works for about 80,000 hours—40 hour work week with 2 weeks annual vacation over 40 years. This would mean the average donor gives the equivalent of 2,000 hours salary. For one hour of research conducted before any donation has been given to yield a negative impact, it would have to have less than a 1/2000 or %0.05 increase in effectiveness.

[3] More precisely, the poorest 19% of Americans live on less than $27.40 a day (US Census Bureau 2013). The poorest 17% of the world's population live on less than $1.50 a day, meaning they are 18 times as destitute (World Bank 2015). Dollar amounts adjusted for purchasing power. Calculations assume income is flat or normally distributed.

[4] Between August 2012 and August 2013, the Make A Wish Foundation of America spent over $246 million (Make a Wish Foundation 'Combined Financial Statements'). In 2014, the foundation granted 14,200 wishes. Assuming expenses for 2014 were approximately equal to 2013, this amounts to $17 thousand per wish (Make A Wish Foundation 'Wish Impact & Facts'). By contrast, donations to the Against Malaria Foundation can save a child's life for $3,340 (GiveWell 2014). This means that the cost of granting a wish is equal to the cost of saving 5 lives.

Why People Donate

Hopefully this evidence is enough to convince us that the overwhelming majority of people are ineffective altruists who behave as if they are mostly indifferent to the effectiveness of their charitable donations. But if people do not donate to charity to minimize suffering, why do they donate to charity? Research in moral psychology has identified two predominant factors—the warm glow of giving and signalling effects. However, while both of these factors influence people to give to charity, they have only a limited ability to influence which charities people give to. As we will see, the cognitive mechanisms responsible for charity choice respond to cues which many of us would consider arbitrary and unimportant.

The warm glow of giving is the subjective feeling of satisfaction we experience when we make a personal sacrifice to help someone else. (see, e.g. Andreoni 1989 and Crumpler & Grossman 2008) We can experience this feeling whether or not we can expect to receive material rewards from our action, suggesting that humans have evolved or acquired an intrinsic motivation to make personal sacrifices for the sake of helping others. This feeling can even be induced when we know ahead of time that our sacrifice will do nothing to further the well-being of the intended recipients. Simply giving is enough to make us feel good about ourselves.

Another reason we give is to show off our moral rectitude. (see, e.g. Lacetera & Macis 2010; Dean & McConnell 2012; and Rand & Nowak 2013) It is important to us that our family, friends, and community members believe we are good people. Giving to charity is one way to demonstrate our altruistic character. This is called a *signalling effect*—when one of the benefits of an action is the signal it communicates to others. In this case, the action is donating to charity, and the signal it sends is that we are kind and caring individuals. As the turn of phrase goes, *be good to seem good*.

These are the two main factors that motivate people to donate to charity. Of course, this psychological evidence does not eliminate the role of helping others as a motivational factor. It is not a coincidence that we experience a warm glow when making a sacrifice *for the sake of helping others,* even when this sacrifice is entirely symbolic, or that the best way to signal we are good people is by doing something *for the sake of helping others*. The evidence simply suggests that helping others is more of an instrumental goal, and holds limited force as an intrinsic motivation.

Charity Choice

For most people, reducing suffering and improving well-being provides little intrinsic motivation to give to charity. But what motivates us to give to certain charities and not others? One third of donors report researching charities before they donate to them, but only 3% report researching cost-effectiveness. (Hope Consulting 2011) Of the donors who do research, only 17% of them aim to find information to compare charities and determine which of them to donate to.(Ibid) And of the donors who do comparison research, just over half of them research cost-effectiveness as a decisive factor. (Ibid) This means that two thirds of individual donors do no research at all, and that 90% of those who do fail to consider cost-effectiveness. So what information *do* we use to decide between charities?

Charity choice for unresearched donations are determined largely by cognitive biases. For example, when we see posters on the metro advertising for a charity you can donate to with via text message, what factors determine whether or not we will do so? Moral psychology has provided us with an extensive list of biases, but I will mention only a few of the most important here:

Physical proximity bias (Musen 2010 as described in Greene 2013)—How far away from me are the recipients of my donation?

Identifiable victim effect (e.g. Loewenstein et. al. 2006)—Do I know any personal information, especially the name and face, of the recipients of my donation?

In-group bias (e.g. Henri & Turner 1979)—Are the recipients of my donation members of my country, or another group I belong to?

These biases may also serve as a heuristic for which charities donors decide to research. For example, imagine a commuter sees one of these advertisements, but never donates to a charity without going on its website. I would conjecture that the commuter is more likely to look up a charity which helps people nearby, shows a picture of an identifiable victim, and works in her own country. When conducting research, a different set of biases come into play, including:

Evaluability bias (Caviola et. al. 2014)—Does the charity score well on easily evaluated measures, particularly low overhead?

Basic- and subordinate- level bias[5]—Does the charity work on a problem that was similar on a basic or subordinate level to a problem that affected me or a loved one?

Even though donors overwhelmingly claim to care about cost-effectiveness when prompted, helping others effectively plays a minimal role in motivating them to donate or determining which charities they donate to. This evidence should lead us to wonder whether ineffective altruism is irrational at all. Perhaps helping others has very little to do with altruism. And perhaps all of these supposed 'biases' we have been discussing are perfectly rational features of our decision-making processes.

I believe such a conclusion would be a mistake. If we had access to better information and took time to reflect on how we choose between charities, I expect most people would realize that what they actually care about is improving the lives of as many people as they possibly can by as much as they possibly can. By contrast, I would wager that most of the factors that currently determine charity choice would seem at best minimally important. The aforementioned psychological mechanisms really are *biases* in the sense that they cause us to behave in ways that we ourselves would disapprove of upon reflection.

We have already seen this revealed preference structure in tests on the evaluability bias.(Caviola et. al. 2014) When asked how much a subject wishes to donate to a charity presented in isolation, subjects' donations correlate more strongly with overhead ratio than cost-effectiveness. However, when subjects are provided with more information and are allowed to compare charities side by side, their donations correlate more strongly with cost-effectiveness than overhead ratio. The conclusion we should draw from this study is that, although people behave as if they care more about overhead than effectiveness, they do so only because of a lack of information. In fact, people care more about helping others effectively, but this preference is only revealed under conditions of better information and reflection. Though the relevant studies have yet to be conducted, I predict there will be similar findings for all of the biases I have just mentioned. To see why, ask yourself about each one in turn:

[5] I hasten to add that I know of no experimental evidence for this bias, so my mention of it here should be taken as speculation based on personal observations about people's motivation for charity choice. I would encourage researchers to explore this bias experimentally.

For example, imagine a woman's child has died of leukemia. There are several levels of abstraction at which she could think about this tragedy, each of which may result in different patterns of charitable giving. She may think, 'I have lost my child to leukemia; therefore I will donate to charities which fight leukemia' (subordinate level), or 'I have lost my child to cancer, therefore I will donate to charities which fight cancer' (basic level), or 'I have lost my child; therefore I will donate to charities which fight the most prevalent causes of child mortality' (superordinate level). However, people tend to focus on the subordinate and basic levels while failing to abstract to the superordinate level.

Physical proximity bias: Does someone's suffering become less important to you as a function of geographic displacement? Would you be willing to pay $100 to save a child's life when she is a mile away from you? What about two, twenty, or one hundred miles? How far away does this child have to be before you would consider it acceptable to let her die for $100?

Identifiable victim effect: Does someone's suffering become less important to you as a function of not knowing her name? What if you determined to donate $100 to save a child whose name you were told but forgot before making the donation? Would this be an acceptable reason to let her die?

In-group bias: Does someone's suffering become less important to you because you happen to have been born in different countries? Would you be willing to donate $100 to save the life of a child from your own country? What if the child moved to a different country? Would this be an acceptable reason to let her die?

Evaluability bias (specifically overhead aversion): Is it worth letting people suffer and die to ensure that the employees and CEOs of a charity get paid less? How much less would a charity's employees and CEOs have to get paid in order for it to be worth letting a child die?

Basic- and subordinate-level bias: Does someone's suffering become less important because they suffer from something that no one you care about has experienced? For example, if a loved one of yours were to die from cancer, would this make children who die from malaria less important than children who die from cancer? Would you be willing to donate $100 to save a child from dying of cancer? How dissimilar does a cause of mortality have to be from cancer in order for you to consider it acceptable to let it kill a child for $100?

When confronted with these sorts of questions, I imagine most people would realize how arbitrary and unimportant factors like physical proximity are to them. By contrast, I predict that cost-effectiveness strikes people as an important factor even when subjected to similar scrutiny.

Effectiveness: Is the suffering of one person less important than the suffering of five people? Would you be willing to pay $100 to save one child's life? If so, does this imply you would be willing to pay more to save the lives of five children? Given the choice between donating to a charity which would use your money to save the life of one child and a charity which would use your money to save the lives of five children, would you choose to save one and let five die, or save five and let one die?

We can subject our biases to the same sort of scrutiny for any type of suffering.

Here I have chosen to focus on child mortality as a prototype cause of misery. But we could equally well ask these questions about, say, rape. For the physical proximity bias we might ask, *How far away does a woman have to be before you would consider it acceptable to allow her to be raped for $100?* My intuition is that it does not matter how far away this woman is—suffering is equally important no matter where it occurs. What *does* matter to me is that I do whatever I can to most effectively mitigate suffering and foster well-being. If you share this intuition, you ought to be an effective altruist as well.

In sum, here is the explanation for why most people are not effective altruists, but should be:

We have psychological incentives to donate to charity, even if only a small part of these incentives is a desire to improve well-being as effectively as possible. While these incentives determine that we should donate to charity, they do not fully specify which charities we should donate to.

Given proper information and rational reflection, we recognize that we would prefer to choose the most effective charities.

However, the psychological mechanisms we currently use to determine our choice of charities rely on factors which, to many of us, seem arbitrary upon reflection.

Therefore, instead of relying on psychological biases, our preferences are better served by choosing charities based mostly if not entirely on effectiveness.

OVERCOMING ALTRUISTIC BIASES

At present, most people give to charity because it gives them a warm glow and a positive reputation, and choose which charities to give to based on cognitive biases. I expect similar psychological mechanisms determine other altruistic decisions, which for some people include volunteering and formulating opinions on how government should litigate for the public good. As a result, there is much more suffering in the world than there would be if only we would act on our altruistic impulses in ways that effectively improved people's lives. Fortunately, there are ways to overcome these biases.

To illustrate my proposal, it will helpful to draw an analogy between how we think about economics and how we think about altruism. Like altruistic decision-making, economic decision-making suffers from a host of cognitive biases. But unlike altruistic decision-making, we have developed methods for recognizing and overcoming biases in economic decision-making. In what follows, I explicate this analogy further and suggest that the methods we employ to think about economics can be used to think about altruism as well.

Overcoming Economic Biases

Consider the life-cycle hypothesis in economics, which holds that individuals prefer smooth consumption throughout the course of their lifetime. (For an early example, see Modigliani 1966) Standard economic theory predicts that, all else being equal, we prefer to consume more rather than less in any given period of time. However, there are diminishing marginal returns on consumption. In any given year, we prefer to consume $75,000 worth of goods to $50,000 worth of goods and $50,000 worth of goods to $25,000 worth of goods, but we more strongly prefer $50,000 to $25,000 than $75,000 to $50,000. Supposing we have a fixed amount of wealth which we can consume at any rate we choose, maximizing utility over the course of our lives requires that we consume at a constant rate.

The extent to which we practice consumption smoothing in real life is constrained by, among other things, psychological biases. We spend impulsively, take on more debt than we can afford, and consistently underestimate how much we need to save for our long-term financial goals. One of the biases that precipitate this behaviour is known as *hyperbolic* temporal discounting.(e.g Madden et. al. 2003 and Green et. al. 1994) Our reflective preferences dictate that we should smooth consumption, but we have an intuitive drive to consume more now and leave less for later. The conflict between immediate and delayed gratification is mediated by two largely independent cognitive processes. (e.g. McClure et. al. 2004 and Metcalfe & Mischel 1999) One—the faster, emotionally charged process—generates a strong, visceral desire to spend now. The other—the slower, emotionally cooler process—implores us to engage in long-term financial planning. Things like saving for retirement require our slower, reasoning processes to direct or perhaps supersede our faster, intuitive processes.

What this means in practice is that we should explicitly recognize our preference for smooth consumption, determine the best way of satisfying this preference using the best epistemic norms available to us, and act according to the conclusions we reach. Many people, for example, hire a financial consultant to help them plan for retirement and attempt to implement her advice by saving and investing accordingly. Not everyone thinks about retirement or relies on epistemically reliable information such as expert advice when doing so. But we all recognize that these are the sort of steps we ought to take if we care about being financially solvent in our later years.

Economic thinking and altruistic thinking have much in common. We are capable of recognizing certain preferences in economic and altruistic decision-making, such as having a smooth consumption curve and donating to effective charities. In both domains, our preferences are hindered by cognitive biases, such as hyperbolic temporal discounting and the physical proximity bias. The conflict between our rationally endorsed preferences and our biases is mediated by similar cognitive processes with similar neural underpinnings. (Greene et. al. 2004; Greene et. al. 2001) It is therefore reasonable to expect that the same mode of thought which allows us to overcome our economic biases can allow us to overcome our altruistic biases as well.

What this involves is a procedure whereby we:

Explicitly recognize our preferences,

Use epistemically reliable methods to decide how best to satisfy these preferences, and

Act on our decisions

In the example of consumption smoothing, we realize that we need to save for retirement, rely on information provided by financial experts, and save and invest accordingly. We can follow a similar process when it comes to altruism. To begin with, we need to recognize our preference for altruistic actions which most effectively improve well-being. The next step is gather information on how best to satisfy this preference. Just as most of us rely on financial experts for advice, the most reliable way to do so—apart from conducting our own extensive research—is to rely on experts such as those at the Centre for Effective Altruism. Finally, we need to implement this advice, perhaps by switching our donations to more effective charities or considering high-impact career options.

Is Effective Altruism Killing the Love?

Before concluding, there is at least one more concern that needs to be addressed. Studies have shown that the employment of reasoning processes in pro-social decision-making tasks correlates negatively with generosity. It is empirically possible, then, that employing reasoning processes in altruistic decision-making will decrease altruism to such an extent that it will more than offset its increase in effectiveness. Paradoxically, it may be more effective to make altruistic decisions based on the very cognitive biases that make our altruism ineffective.

This is an interesting possibility, but empirically implausible. Though no studies have tested this directly, related research shows that employing reasoning processes under certain conditions can decrease altruism by 15-50%[6]. But considering some charities are thousands of times more effective than others—for example, with donations to guide dog charities versus trachoma charities—it would be surprising to learn that rational thinking increases the effectiveness of our giving by less than a factor of two. On empirical grounds, the expected increase in effectiveness eclipses the expected decrease in altruism. I would also speculate that the sort of people who engage in rational thought for the express purpose of helping others as much as they possibly can will be among the least susceptible to having their motivation desiccated by reasoning processes. Perhaps the tradeoff between effectiveness and altruism is not such a problem outside the lab. While this is still an open question, the available evidence suggests that rational thought is essential for effective altruism.

CONCLUSION

Today, most people are ineffective altruists. We perform actions for the sake of helping others, but we do so in such a way that gives less help to fewer people than we otherwise could. Most of our motivation for donating to charity comes from a desire to feel good about ourselves and score reputation points. And most of what determines our choice of which charities to donate to is a collection of cognitive biases. As a result, millions of people and non-human animals will continue to suffer unnecessarily.

Effective altruism is the antidote to this miserable state of affairs. Concisely put, effective altruism is about "aiming to do the most good that one can". I have offered a more precise explanation of what this means, and shown it to be a much milder and more intuitive philosophy than utilitarianism. We do not have to accept the re-

[6] Rand et. al. 2012 shows a 15% decrease in contribution to public goods games; Loewenstein et. al. 2006 shows a 50% decrease in charitable contribution to a statistical victim versus an identifiable victim.

pugnant conclusion or consider ourselves morally blameworthy for actions which are not bad for anyone in order to be effective altruists. Nor do we have to relegate considerations of deontological values like justice and fairness to a role of merely instrumental importance. All we have to believe is that when we act altruistically, it is preferable to give more help to more people, rather than less help to fewer people, all else being equal.

If we are to live in accordance with this preference, we need to revolutionize how we think about altruism. In addition to thinking intuitively, we need to think rationally. I suggest that we reconceive of altruism in economic terms, whereby we view acts of charity as an investment in the well-being of valuable creatures. And we should demand nothing less of ourselves than to see our investment yield maximum returns. Making even the simple decision to donate to effective charities can increase our impact by orders of magnitude. Faced with these facts, it should be evident by the light of our own values that it is no longer acceptable to just make the world a *better* place. This is too modest a goal. Instead, we should endeavor to improve the lives of as many people as possible by as much as possible, and use our altruism to do the most good we can.

Acknowledgements: I would like to thank the Oxford Uehiro Centre for organizing this essay contest, my friends from Crockett Lab - especially Lucius Caviola, Andreas Kappes, and Molly Crockett - for their enormous influence on my thinking about moral psychology, my friends from the Centre for Effective Altruism - especially William MacAskill and Hauke Hillebrandt - for introducing me to a wonderful new way of thinking about moral philosophy, and my advisor Professor Daniel Dennett to whom I owe a greater debt of intellectual gratitude than I could ever hope to repay.

REFERENCES

Andreoni, J. (1989). Giving with impure altruism: applications to charity and Ricardian equivalence. *The Journal of Political Economy*, 1447-1458.

Caviola, L., Faulmüller, N., Everett, J. A., Savulescu, J., & Kahane, G. (2014). The evaluability bias in charitable giving: Saving administration costs or saving lives?. *Judgment and Decision Making*, 9(4), 303.

Crumpler, H., & Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of Public Economics,* 92(5), 1011-1021.

Gallup. (2013, December 13). Most Americans Practice Charitable Giving, Volunteerism. Retrieved June 15, 2016, from http://www.gallup.com/poll/166250/americans-practice-charitable-giving-volunteerism.aspx .

GiveWell. (2014, November). Against Malaria Foundation (AMF) | *GiveWell*. Retrieved July 24, 2015, from http://www.givewell.org/international/top-charities/amf.

Giving USA. (2015, June 29). Giving USA: Americans Donated an Estimated $358.38 Billion to Charity in 2014; Highest Total in Report's 60-year History. Retrieved June 15, 2016, from http://givingusa.org/giving-usa-2015-press-release-giving-usa-americans-donated-an-estimated-358-38-billion-to-charity-in-2014-highest-total-in-reports-60-year-history/.

Green, L., Fry, A. F., & Myerson, J. (1994). Discounting of delayed rewards: A life-span comparison. *Psychological Science,* 5(1), 33-36.

Greene, J. D., Nystrom, L. E., Engell, A. D., Darley, J. M., & Cohen, J. D. (2004). The neural bases of cognitive conflict and control in moral judgment.*Neuron*, 44(2), 389-400.

Greene, J. D., Sommerville, R. B., Nystrom, L. E., Darley, J. M., & Cohen, J. D. (2001). An fMRI investigation of emotional engagement in moral judgment. *Science*, 293(5537), 2105-2108.

Greene, J. D. (2013). *Moral tribes: Emotion, Reason, and the Gap Between Us and Them*. New York: Penguin.

Guide Dogs of America. (2015). Donating FAQ. Retrieved June 15, 2016, from http://www.guidedogsofamerica.org/1/help/donate/.

Hope Consulting. (2011, November). Money for Good II: Driving Dollars to the Highest Performing Non-Profits. Retrieved June 15, 2016, from https://www.guidestar.org/ViewCmsFile.aspx?ContentID=4038.

Karlan, D., & McConnell, M. A. (2014). Hey look at me: The effect of giving circles on giving. *Journal of Economic Behavior & Organization*, 106, 402-412.

Lacetera, N., & Macis, M. (2010). Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2), 225-237.

Loewenstein, G., & Deborah, A. Small, and Jeff Strnad. 2006. Statistical, Identifiable, and Iconic Victims. *Behavioral Public Finance*, 32-46.

Madden, G. J., Begotka, A. M., Raiff, B. R., & Kastern, L. L. (2003). Delay discounting of real and hypothetical rewards. *Experimental and Clinical Psychopharmacology*, 11(2), 139.

Make a Wish Foundation. (2014, February 04). Combined Financial Statements. Retrieved July 24, 2015, from http://wish.org/~/media/100-000/About%20Us/Making%20a%20Difference/Managing%20Our%20Funds/Documents/FY2013/FY13%20MAWFA%20Combined%20FS_Final%202_04_14.ashx?la=en.

Make-A-Wish America. (2011). Wish Impact & Facts. Retrieved June 15, 2016, from http://wish.org/wishes/wish-impact#sm.00000r9h9eqoladpqvbz65boolroh.

McClure, S. M., Laibson, D. I., Loewenstein, G., & Cohen, J. D. (2004). Separate neural systems value immediate and delayed monetary rewards.*Science*, 306(5695), 503-507.

Metcalfe, J., & Mischel, W. (1999). A hot/cool-system analysis of delay of gratification: dynamics of willpower. *Psychological Review*, 106(1), 3.

Modigliani, F. (1966). The life cycle hypothesis of saving, the demand for wealth and the supply of capital. *Social Research,* 160-217.

Ord, T. (2013, March). The moral imperative toward cost-effectiveness in global health. *Center for Global Development,* 1-12. Retrieved June 15, 2016, from http://www.cgdev.org/sites/default/files/1427016_file_moral_imperative_cost_effectiveness.pdf.

Raihani, N. J., & Smith, S. (2015). Competitive helping in online giving.*Current Biology*, 25(9), 1183-1186.

Rand, D. G., & Nowak, M. A. (2013). Human cooperation. *Trends in Cognitive Sciences,* 17(8), 413-425.

Rand, D. G., Greene, J. D., & Nowak, M. A. (2012). Spontaneous giving and calculated greed. *Nature*, 489(7416), 427-430.

Singer, P., & MacAskill, Will. (2015). Introduction. In R. Carey, *Effective Altruism Handbook*. (viii-xvii). Oxford: Centre for Effective Altruism.

Singer, P. (1972). Famine, affluence, and morality. *Philosophy & Public Affairs*, 229-243.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict.*The Social Psychology of Intergroup Relations*, 33(47), 74.

US Census Bureau. (2014). Selected Income Characteristics. Retrieved June 15, 2016, from http://factfinder.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=ACS_12_5YR_DP03.

World Bank. (2015). Poverty Overview. Retrieved July 24, 2015, from http://www.worldbank.org/en/topic/poverty/overview.